Simplicial Centrality Measure Applications: Protein-Protein Interaction and Wireless Sensor Networks

Grant J Ross

Applied Analysis

Department of Mathematics and Statistics

University of Strathclyde, Glasgow

August 7, 2025

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: Grant Jamieson Ross

Date: 11 October 2024

Abstract

Using networks to model discrete systems of interactions is very common in the literature. However, networks are limited to considering only paired interactions. One of the responses to this limitation is to extend the concept of an edge to contain more than two nodes. Simplicial complexes are one of the models of discrete interactions which features this extension and we shall study them here. Centrality measures are one of the main ways of studying networks and are used to discover which node is most important. We shall extend centrality measures to the case of simplicial complexes. We shall then use these centrality measures on simplicial complexes to analyse protein–protein interaction networks and detect nodes whose removal would cause a gap in coverage on wireless sensor networks. We also assess the ways that dynamic systems work on random geometric graphs which have central sections missing.

Contents

List of Figures

List of Figures

List of Tables

List of Tables

Preface/Acknowledgements

I would not have been able to complete this thesis without the support of several amazing and wonderful people. I would like to start by thanking my supervisor Dr Matthias Langer for his patience, excellent advice and guidance and always being pleasant to chat to. I would also like to thank the EPSRC for funding my studentship through grant EP/M508159/1. My senior friend Dr Alison Ramage and Irene Spencer also provided great support and help that I would have been lost without.

I would like to thank Barry Herron and Alan Paget, my line managers, for approving my Monday off without which I wouldn't have had the time to complete this thesis.

I would also like to thank my friends for encouraging me and providing welcome distraction to the stresses of putting this work together. Particular thanks go to Patrick Campbell and Sium Ghebru for always checking in and encouraging me and to Luke Delello for that and also for the solidarity while we worked on our projects together.

The Dolmaire family have made me feel so welcome and are always very patient with me.

I would like to thank Larry and Louise Jamieson for the encouragement, the advice, the laughs and the rhubarb.

I would like to thank my Granny for teaching me that "if at first you don't succeed try try again" and my Granda for encouraging me to ask questions and keep an open mind.

I would like to thank my sister Ailsa and her husband Jeremy Scott for engaging in japes and not letting me take myself too seriously.

Massive thanks to my delightful parents Len and Caroline Ross for providing me with the encouragement and the platform to learn and grow into the person I am today.

Also for always being there to offer advice in any situation, being unconditional in your love and support for me and all our lovely conversations. Also thank you for reminding me to keep going, that it will be worth it in the end and generally annoying me into finishing this work.

I would like to thank my wonderful partner Emma Dolmaire who has given me so much kindness, love and encouragement and has allowed me to focus on getting this thesis done. You are wonderful.

Chapter 1

Introduction

Networks have been employed in the study of a wide variety of scientific disciplines since their introduction by Leonhard Euler to solve the seven bridges of Königsberg [53] (in English [54]). They have been used to study transport links [32], neurons in the brain [12] and food webs [37] among many others [112]. In Chapter 2 we introduce network-theoretic concepts as well as the other theorems and results which underpin the work we have done.

However, networks have limitations. Consider a social network with three individuals where they are connected if the individuals speak to each other on a given day. If all three nodes are connected then this situation could have come about because each pair of individuals spoke to each other once throughout the day or due to a single three way conversation. There is no way to distinguish between these situations in a network-theoretic environment.

One way of addressing this problem is through the use of simplicial complexes which have also been studied a lot in the literature [66, 71, 97, 107, 108, 147, 153]. Simplicial complexes allow connections between three or more nodes but are also closed in the sense that a connection between three nodes also implies a connection between each pair of nodes. In Chapter 3 we expand some network-theoretic concepts to the case of simplicial complexes. Our method is based on an extension of adjacency to isolate the different dimensional levels and from there we generalise walks and centrality measures to the simplicial complexes case.

We apply these centrality measures to solve network-theoretic problems. In Chapter 4 we apply them to protein–protein interaction (PPI) networks, which are networks based on interactions between proteins in the cells of all species of plants, animals, bacteria and viruses among others [168]. In these networks proteins come together to produce many of the reactions which allow the cells to live. These interactions frequently involve multiple different proteins, which makes them a prime candidate to be looked at through a simplicial complexes lens.

We use a couple of different approaches to analyse protein–protein interaction simplicial complexes. First we look at the degree distributions across a variety of different PPIs and demonstrate that although the one and two dimensional levels of PPI simplicial complexes have some high degree simplices there are many simplices of medium degree which contrasts with the situation for PPI networks where a small number of hub nodes have very high degree and the majority of the rest of the nodes have much lower degree. Secondly, we identify essential proteins which are proteins which, when removed, cause the death of the cell. Network-theoretic centrality measures have previously been used to search for these proteins with fair success under the centrality-lethality paradigm [73]. We use the centrality measures for triangles and show that they can identify more essential proteins than their node-based counterparts and that they also identify different ones, which suggests the two approaches could be combined to maximise the number of essential proteins found.

In Chapter 5 we move from the triangular centralities to the edge-based ones. We apply them to wireless sensor networks (WSNs) which are networks of sensors which are monitoring and/or communicating with each other usually over a remote or inhospitable area [10, 82]. These WSNs are usually modelled as random geometric graphs (RGGs) [63]. Gaps in the area covered by these networks can be problematic because they could result in missed data which may lead to problems in predicting whether or not an area is safe or missing a problem which needs to be addressed. As a result detecting gaps in the coverage of these networks has been a popular topic of study [34, 126]. However, as Kenniche and Ravelomananana [82] note, there are wireless sensor networks where the sensors are fitted with batteries which it may not be possible to recharge or replace.

Sensors may also fail for other reasons and it may not be possible to replace them quickly, for example a tsunami detection sensor in a remote part of the ocean. For this reason we should not limit our interest to the detection of holes, we should also seek to create robustness within these networks which means that we want to be able to detect nodes whose removal would create a hole. We demonstrate that edges which are longer in an RGG are expected to have larger edge degrees and use this fact to demonstrate that nodes which are members of edges which have high edge subgraph centrality are likely to be isolated from other nodes and so are more likely to cause a hole to appear in the event of their removal. We finish this chapter by experimentally demonstrating the effect of this strategy for removal of sensors compared to three other methods.

In Chapter 6 we move away from simplicial complexes to consider another application of RGGs, the spread of plant disease through a region. Historically, RGGs have been studied on square areas but recently other shapes have been considered. Sheerin and Estrada looked at random rectangular graphs in 2015 [51] and Giles et al. considered annuli in 2016 [64]. We combine these two situations and look at random rectangular annular graphs (RRAGs). We derive a formula for the expected mean degree of these constructions. We then demonstrate that elongation of the annular areas and having many small annular areas rather than one large one is likely to decrease the mean degree. The epidemic threshold of a disease spreading on a graph is bounded by the reciprocal of its mean degree [112]. This connection suggests that an RRAG constructed on a region with many small annular areas where nodes cannot be spread will have a larger epidemic threshold than an RRAG constructed with one large such area. We show that there is evidence of this effect through the use of simulations.

Chapter 2

Preliminaries

In this chapter we introduce much of the work that has previously been done that we rely on in later chapters. We look at: network theory, what a network is and what processes we can model with them; centrality measures and how we can decide which parts of a network are most important; Laplacian matrices and how they model flow of quantities through the system and; simplicial complexes which we use to demonstrate the value of looking beyond pairwise interactions.

Parts of this chapter formed preliminary sections of [50].

2.1 Network Theory

2.1.1 Nodes, Edges and Adjacency: When Are Two Things Connected?

The first instance of a graph theoretic proof was Leonhard Euler's tackling of the problem of the seven bridges of Königsberg [53,54]. The problem, of endless fascination to the people of the 18th century Prussian city, was to find a way to navigate around the city and cross each of its seven bridges precisely once. His leap in terms of proving that such a walk was impossible was to consider each of the four landmasses which made up the city as a separate identity, assigned to a letter, and the bridges as connections between the landmasses.

Network Theory is a branch of Graph Theory which uses graphs to model phenom-

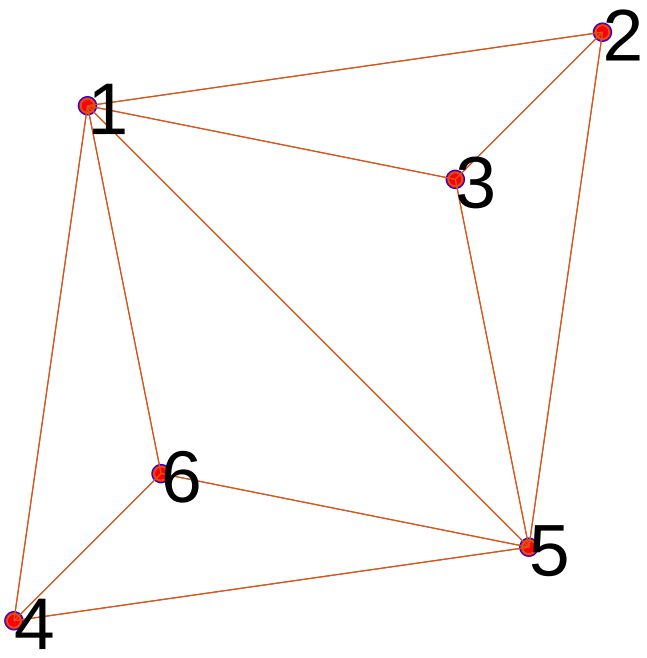


Figure 2.1: A network with labelled nodes.

ena from other scientific disciplines.

Definition 2.1. A **network** or **graph** is a pair $G = (V, E)$ which consists of a set $V = \{v_1, v_2, ...v_n\}$ of n **nodes** and another set E of two-element subsets of those nodes called the **edges** [46, Defn 2.1].

For example, in the network depicted in Figure 2.1 the set of nodes is given by $V = \{1, 2, 3, 4, 5, 6\}$ and the set of edges is given by $E = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5\}, \{1, 6\}, \{2, 3\}, \{2, 5\}, \{3, 5\}, \{4, 5\}, \{4, 6\}, \{5, 6\}\}$.

It should be noted that the definition of a network given here is quite restrictive in the sense that it does not permit double edges (the same pair of nodes forming two edges) or loops (an edge which contains the same node twice). It also excludes directed graphs, which are commonly used where there is a sense of flow in a network. This choice was a deliberate decision on the part of the author to prioritise simplicity because none of the applications which have been developed in subsequent chapters rely on any of these mechanisms.

It is possible to create new networks by considering only a certain part of an existing

network.

Definition 2.2. A graph $G_s = (V_s, E_s)$ is a **subgraph** of a graph $G = (V, E)$ if $V_s \subseteq V$ and $E_s \subseteq V_s \times V_s \cap E$ [46, Defn 2.2].

We illustrate this concept with some examples in reference to the network in Figure 2.1. If we take the nodes $V_s = \{1, 3, 5, 6\}$ to be the set of nodes of our subgraph then we have a valid subgraph if $E_s = \{\{1, 3\}, \{1, 5\}, \{1, 6\}, \{3, 5\}, \{5, 6\}\}$ which is all of the edges between nodes in V_s. Additionally, we also have a valid subgraph if $E_s = \{\{1, 3\}, \{5, 6\}\}$ because it is not necessary for all possible members of $V_s \times V_s \cap E$ to be members of E_s. However, we cannot add new edges that are not present in the original graph nor can we include edges which contain nodes which are not members of V_s so $\{3, 6\}, \{4, 6\} \notin E_s$.

When one is using networks to model a phenomenon from another scientific discipline it is best practice that there is a clear and natural understanding of what the nodes should be and what the edges represent. In Chemistry [24] networks have been used to analyse the interactions of proteins in cells by taking a node to represent a single protein and an edge exists between two nodes if their proteins are involved in the same cellular process. We discuss protein–protein interaction networks further in Chapter 4. In an airport network [32, 68] each node is an airport and there is an edge between two nodes if a regular direct flight exists between those locations. Further examples of phenomena which have been modelled using networks include: networks of neurons in the brain [12] where the nodes are individual or groups of neurons and the edges represent whether or not the nodes activate at the same time; disease transmission networks [145] where nodes represent individual entities and there exists an edge between two nodes if it is possible for a disease to transmit from one to the other; and food webs [37] where the trophic relations (which species eat which other species) between animals in an ecosystem are studied through network theory, with each species assigned to its own node and with edges going from predator to prey species (this use case is an example of a directed network). There are many other areas of science which have been modelled using networks [112].

Consider the network in Figure 2.1, the 6 nodes are the characters in the sitcom

Friends and there exists an edge between two of the nodes if they ever lived together during the show. Using the encoding that node 1 represents Chandler, 2 Phoebe, 3 Monica, 4 Joey, 5 Rachel and 6 Ross we can represent this system by that network.

Definition 2.3. Two nodes $v_i, v_j \in V$ are **adjacent** if there is an edge $e = \{v_i, v_j\} \in E$ between them [46, Defns 2.2 and 2.3]. The **adjacency matrix** of a network is the matrix defined as follows:

Let $v_i, v_j \in V$ be two nodes in a network. Then, the adjacency matrix A has entries defined by

$$A_{ij} = \begin{cases} 1 & \text{if } v_i \text{ is adjacent to } v_j; \\ 0 & \text{if } v_i \text{ is not adjacent to } v_j \text{ or } i = j; \end{cases}$$

We now have a mathematical representation of a network which we can perform calculations on to discover facts about the network. As an example the network from Figure 2.1 can be represented by the matrix below, where the rows and columns representing the nodes are arranged numerically:

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

On certain kinds of networks, the Königsberg bridges being a prime example, a natural question is: What does it look like to move around the network? To that end we have the following definition.

Definition 2.4. A **walk** on a network is a sequence of (not necessarily distinct) edges $\{u_1, v_1\}, \{u_2, v_2\}, ..., \{u_k, v_k\}$ for which $v_i = u_{i+1}$ $(i = 1, 2, ..., k-1)$. If $v_k = u_1$ then the walk is **closed**. A **trail** is a walk in which all the edges are distinct. A **path** is a walk where the u_i are distinct. A **shortest-path** between two nodes, n_i, n_j, is a path such that $u_1 = n_i$ and $v_k = n_j$ which minimises number of edges traversed. The **shortest-path distance** between two nodes is a function, d, on the nodes of a network, V, such

that $d : V \times V \to \mathbb{R} \cup \{\infty\}$, where the result, $d(v_i, v_j)$ is given by the number of edges in the shortest path between v_i and v_j. If there is no path between v_i and v_j then $d(v_i, v_j) = \infty$. A **cycle** is a closed path. ([46, Defn 2.5] and [43, p. 47])

It is well known within network theory that by taking powers of the adjacency matrix of a network the entries of the resulting matrix give the number of walks of that length between the two relevant nodes [36, Thm 0.1]. Therefore the number of cycles of length p starting and ending at node i is given by $(A^p)_{i,i}$.

For example, in the network shown in Figure 2.1

$$\{6,5\},\{5,1\},\{1,4\},\{4,5\},\{5,2\} \tag{2.1}$$

is a walk but not a path because the node 5 is repeated, whereas

$$\{6,5\},\{5,1\},\{1,4\} \tag{2.2}$$

is a path.

An example of a cycle is

$$\{6,5\},\{5,1\},\{1,4\},\{4,6\}. \tag{2.3}$$

There are two shortest paths between node 6 and node 2 which are

$$\{6,5\},\{5,2\} \tag{2.4}$$

and

$$\{6,1\},\{1,2\}. \tag{2.5}$$

The edge $\{6,2\}$ does not exist and so there is no path of length 1. Therefore the shortest-path distance between node 6 and node 2 is 2.

We can use the concept of the shortest-path distance to define an extended metric on a network.

Definition 2.5. A **extended metric**, d, is a function on a set, Q, such that $d : Q \times Q \to$

$\mathbb{R} \cup \{\infty\}$ which satisfies the following axioms for all $a, b, c \in Q$

1. The distance from a point to itself is 0, i.e. $d(a, a) = 0$.
2. The distance between two separate points is always greater than 0, i.e.

   $d(a, b) = 0 \iff a = b$.
3. The distance from a to b is equal to the distance from b to a, i.e. $d(a, b) = d(b, a)$.
4. The triangle inequality is true, i.e. $d(a, c) \leq d(a, b) + d(b, c)$.

[13, p. 128]

Theorem 2.6. *The shortest-path distance is an extended metric on the set V of nodes of a network [43, p. 47].*

The shortest-path distance between two nodes in a network is not always finite. This situation would happen in the case that the network was disconnected for example.

Definition 2.7. A network is **disconnected** if there exists at least two nodes $v_i, v_j \in V$ such that $d(v_i, v_j) = \infty$, that is there does not exist a path between v_i and v_j. A network is considered **connected** if it is not disconnected. A **connected component** of a network is a subset, S of its nodes such that for all $v_i, v_j \in S$ then $d(v_i, v_j)$ is finite and for any $v_i \in S$ and $v_k \in V \setminus S$ then $d(v_i, v_k) = \infty$. [43, p. 18 and p. 47]

Note that the networks in Figures 2.1 and 2.2 are both connected but we can also have a network which combines these networks together as in Figure 2.3 by relabelling each of the nodes $i \in V_2$ to be $i + 6 \in V_1 \cup V_2$ where V_1 is the set of nodes from Figure 2.1 and V_2 is the set of nodes from Figure 2.2. We can represent the adjacency matrix of this expanded network in the form,

$$\begin{pmatrix} A_1 & \mathbf{0} \\ \mathbf{0} & A_2 \end{pmatrix}$$

where A_1, A_2 are the adjacency matrices of the networks with node sets V_1, V_2 respectively and $\mathbf{0}$ signifies a matrix of zeros of the appropriate size. This matrix is said to be in block-diagonal form.

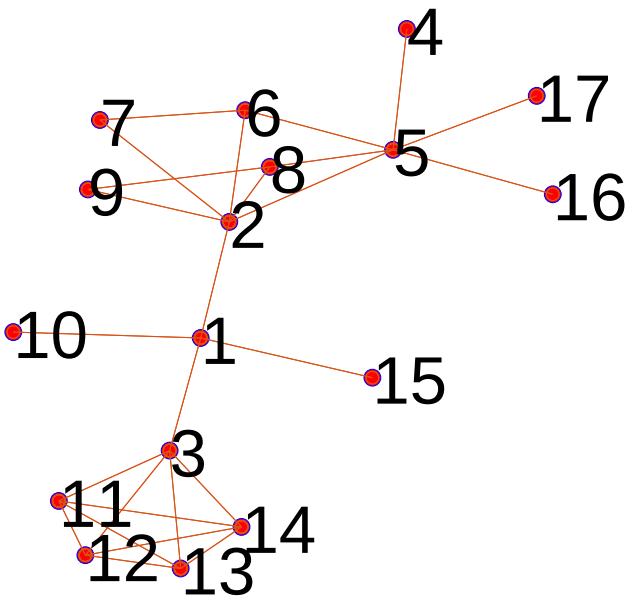


Figure 2.2: A network with labelled nodes.

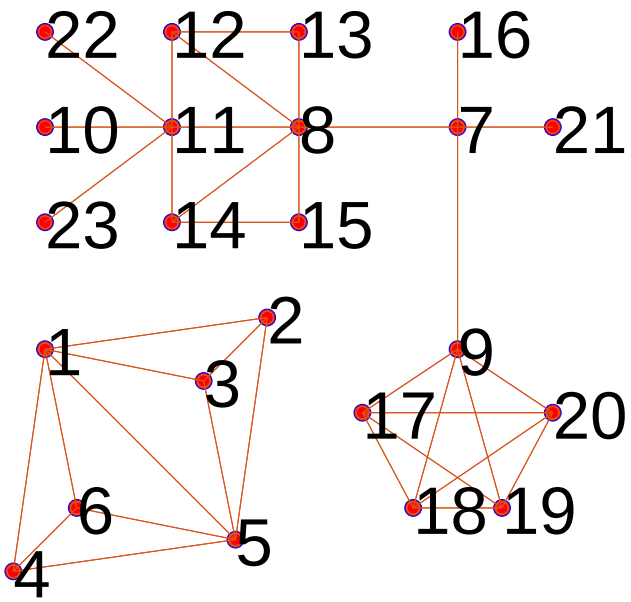


Figure 2.3: A disconnected network.

Definition 2.8. A matrix is in **block-diagonal form** when all non-zero elements are in square blocks along the diagonal of the matrix and all other elements are 0. [112, Section 6.11]

Any network with k connected components can have its adjacency matrix represented in block diagonal form with k blocks by appropriate relabelling in the same manner as above and would be represented as:

$$\begin{bmatrix} A_1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & A_2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_3 & \ddots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & A_k \end{bmatrix}. \tag{2.6}$$

There are some graphs which belong to special families. One such family will be needed for constructing simplicial complexes in Section 2.5.

Definition 2.9. A **complete network** of size n is the network with n nodes where each node is adjacent to every other node [46, p. 15].

A **clique** in a network, $G = (V, E)$, is a subset of its nodes $V_s \subseteq V$ such that it is possible for a subgraph $G_s = (V_s, E_s)$ of G to be a complete network [46, p. 25].

In the network depicted in Figure 2.2 the nodes $\{3, 11, 12, 13, 14\}$ form a clique of size 5 because it is possible that in a subgraph based on these nodes every node would be adjacent to every other node. Similarly, we can see that $\{2, 5, 8\}$ is a clique of size 3 but $\{2, 5, 6, 8\}$ is not a clique of size 4 because there is no edge $\{6, 8\}$.

2.1.2 Linear Algebra

Clearly, given that a network can be represented in matrix form, the study of matrices and associated vector spaces is an important part of network theory. An undergraduate level of understanding of fields and vectors is assumed which includes the definitions of these two concepts as well as an appreciation of linear independence and spanning sets

in vector spaces. We also assume an understanding of how to perform addition, scalar multiplication and matrix multiplication on matrices, how to calculate the determinant of a matrix and a knowledge of how to solve systems of linear equations by using Gaussian elimination to reduce the augmented matrix representation of such a system into reduced row-echelon form.

Nair and Singh [111] provide a good introduction to any of these concepts.

There are many applications for eigenvalues and eigenvectors of both the adjacency matrix and the Laplacian matrix which is introduced in Section 2.3.

Definition 2.10. An **eigenvector** of a matrix, M, is any non-zero vector, p, such that the only action of the matrix on the vector is to multiply it by a scalar, λ [111, Defn 5.1]. That is, p is an eigenvector if

$$Mp = \lambda p. \tag{2.7}$$

The scalar, λ, is known as an **eigenvalue** [111, Defn 5.1].

The **characteristic polynomial** of a matrix, A, is the result of $\det(M - \lambda I)$, where I is the identity matrix [111, Defn 5.7]. The roots of this polynomial are also the eigenvalues of A.

The **algebraic multiplicity** of an eigenvalue is the number of times it is a root of the characteristic polynomial. It can also be thought of as the maximum possible number of linearly independent eigenvectors of a matrix which correspond to that eigenvalue [111, Defn 5.8].

The **geometric multiplicity** of an eigenvalue is the number of linearly independent eigenvectors of a matrix which correspond to that eigenvalue [111, Defn 5.14].

In Section 2.6.3 we will introduce Hodge Laplacian matrices of simplicial complexes. The algebraic/geometric multiplicity of the 0-eigenvalue of these Hodge Laplacians gives us information about the topology of the simplicial complex. We will use these multiplicities in Chapter 5 to detect holes in Vietoris-Rips complexes which have been constructed from Random Geometric Graphs.

2.2 Centrality Measures: What Makes A Node Important?

Consider the worldwide airport transportation network analysed by Guimerà et al. [68]. They wished to figure out whether the cities which had the most connections to other cities were also the cities positioned on the most shortest paths between two other cities. They found that there were some cities like Paris, New York and Frankfurt which had many connections to other cities and also sat on many shortest-paths. However, there were only ten cities which were in the top 25 of both classifications. Cities like Atlanta, USA were connected to a large number of other cities, at the time the network was constructed, but were not in the top 25 airports in terms of lying on the most shortest paths. It was cities like Anchorage, USA and Port Moresby, Papua New Guinea that replaced Atlanta in the top 25 based on the shortest-path classification.

Guimerà et al. reasoned that the reason for this difference between the different classifications is due to Anchorage being the only city in Alaska which has connections to cities outside of Alaska. This change is important because although the closure of the airports in Atlanta would be expected to have a large general disruption to the most people the closure of the airports in Anchorage would prevent anybody in Alaska from being able to leave Alaska by air which demonstrates that the answer to the question which airport is most important depends on the perspective of the person being asked.

Definition 2.11. A **centrality measure** is a function, $f : V \to \mathbb{R}$, from the nodes of a network to the real numbers. The higher the number a node is assigned by a centrality measure the more **central** or important the node is considered to be [140, p. 11].

There are a great number of different ways the notion of importance can be defined; thus there are also many different centrality measures. They are used by Google to assess which pages to recommend to people using their search engine [23], have been combined to investigate the most important patches of habitat to protect in order to minimise ecosystem damage [44] and topological centrality can be used to discover communities in research co-authorship networks [174].

Schoch [140] constructs an illustrative table which classifies the various centrality

measures into several groups. His work also points to the abundance of such measures. Indeed it is important to be very careful when introducing new centrality measures that one is not just adding to the existing noise. For example, there are eight variants of betweenness centrality that differ in weighting factors, the kind of path being considered and the positioning on the path [140]. In this work we shall endeavour, where new centrality measures are introduced, to demonstrate why they are useful.

Below we present some of the more common centrality measures used within network theory.

The simplest of all centrality measures is the degree.

Definition 2.12. The **degree** of a node n is the number of edges of which n is a member [46, p. 143].

For example, the degree centralities of the nodes from Figure 2.1 are 5 for node 1 and node 5 and 3 for the other four nodes. In the network from Figure 2.2 node 5 has degree 6 and node 12 has degree 4.

Having established the concept of the degree of a node we can start comparing them not only within the network but also between networks. The degree distribution of a network describes what proportion of nodes in a network have high degree, what proportion have low degree and how those proportions change between the two extremes.

Definition 2.13. Let $p(k)$ be the probability of finding a node of degree k in a network, then the **degree distribution** of the network is the probability distribution of the degrees of the nodes across the whole of the network [46, p. 95 - 98].

A network with a long tailed degree distribution has a few hubs, high degree nodes, which are well connected and many low degree nodes which are not. These hubs are important from the structural and functional point of view in these networks although lower degree nodes can also be very important in the network such as Anchorage in the earlier airport network example [68]. In later chapters the concept of a degree distribution shall be extended to the case of simplicial complexes and applied to protein–protein interaction networks.

In the Friends example (see Figure 2.1) we have that $p(3) = 0.667$ and $p(5) = 0.333$. Whereas, in the network from Figure 2.2 $p(4) = \frac{5}{17}$ because nodes $(1, 11, 12, 13, 14)$ all have degree 4.

We can also connect the degree to the largest eigenvalue of the adjacency matrix through a well known theorem [112, p. 699].

Theorem 2.14. *The largest eigenvalue of the adjacency matrix, λ_1, is always greater than or equal to the average degree, $\overline{k}$. That is*

$$\lambda_1 \geq \overline{k}. \tag{2.8}$$

We will use this theorem to obtain a bound on the epidemic threshold (to be introduced in Section 2.4) of a Random Rectangular Annular Graph (RRAG), via our calculation of the expected average degree of such graphs in Chapter 6. These calculations will give us a theoretical insight into the requirements for a disease to become an epidemic on RRAGs.

2.2.1 Closeness Centrality

Closeness centrality is a concept which was first introduced by Bavelas [14] to capture the idea of how close, in terms of shortest path distance, a node is to all the other nodes in a network. Later we generalize this concept to simplicial complexes and use it to study protein–protein interaction networks and the effects of removing nodes from random geometric simplicial complexes.

Definition 2.15. The **farness** of a node n is the sum of its shortest path distances to all other nodes, $\sum_{a \neq n} d(a, n)$ [112, p. 181].

Definition 2.16. The **closeness** is the reciprocal of farness [112, p. 182]. That is:

$$C(n) = \frac{1}{\sum_{a \neq n} d(a, n)}. \tag{2.9}$$

It is also common to normalise the closeness centrality by multiplying the results by $(n-1)$. This normalisation means that in any network a node that is adjacent to all

the other nodes of a network has a normalised closeness centrality of 1, which allows for comparisons between networks. Where values for closeness centrality are reported in this thesis they are normalised values.

Note that if the network is not connected then $\sum_{a \neq n} d(a, n)$ could be considered undefined or ∞ for all nodes in the network. In this case we can calculate harmonic closeness [130] instead.

Definition 2.17. The **harmonic closeness** of a node n is defined as follows

$$H(n) = \sum_{a \neq n} \frac{1}{d(a, n)}, \tag{2.10}$$

where we treat $\frac{1}{\infty} = 0$ [112, p. 184].

Closeness centrality has been used in applications as diverse as mapping networks of terrorist cells [89] and the density of commercial activity in cities [122].

For the network in Figure 2.1 nodes 1 and 5 both have a closeness value of 1 because they are both adjacent to all of the other nodes in the network. In Figure 2.2 the top six nodes are node 2 at 0.533, node 1 at 0.5161, node 5 at 0.4324, node 3 at 0.4211 and nodes 6 and 8 at 0.4. We see that this centrality measure prioritises the locations where the fewest steps are required to get to the most locations. If we imagine that the nodes represent airports and the edges connections between those airports then a frequent traveller may want to use the closeness centrality to decide where best to live. In this case node 2 benefits from having the most direct connections compared to node 1 even though it requires 3 trips to get to 4 nodes as opposed to 3 trips for node 1.

2.2.2 Subgraph Centrality

The Closeness Centrality only considers the shortest way of getting between two nodes which is not a realistic way to look at all phenomena which can be modelled by networks. Consider a social network where the nodes are people in an office and the edges represent whether or not they interact on a given day. If we wanted to track which nodes were most influential in spreading a piece of information or a disease around the network then it is not enough to think about how many other nodes a node is connected to. For

example if a node has eight neighbours, of which seven are not connected to any other nodes except the original then the only way from this node to the rest of the graph is through its eighth neighbour. This node is very influential in spreading the information to its eight neighbours but it is not very influential within the rest of the network. The more nodes there are in the rest of the network the more we can say that this node is not very important within the network. And yet this node will likely score highly for degree centrality. Similarly, there is a good chance that the node will have high closeness centrality as a result of its length 1 shortest path to each of its neighbours.

Similarly, a disease or piece of information does not know which path is the shortest to get through a network and is unlikely to be targeting a specific person. Therefore, if there are multiple paths to get from one node to another then we would want to include all of them although we would want to give more weight to shorter walks than others. This desire leads us to the branch of centrality measures which rely on taking powers of the adjacency matrix.

Historically, for networks the first of these centralities was developed by Katz [81]. The Katz centrality index is $\left[\sum_{m=0}^{\infty}(\alpha^m A_k^m)\mathbf{1}\right]_i$ where $\mathbf{1}$ is a vector of 1s. This calculation boils down to summing up the number of walks of each length which originate at a node with longer walks penalised more strongly by ensuring $0 \leq \alpha < \frac{1}{\lambda_1(A)}$ where λ_1 is the largest eigenvalue of A. The eigenvector centrality can be derived from the Katz centrality. The eigenvector centrality of a node, i, is the ith entry of the eigenvector associated with the largest eigenvalue of the adjacency matrix. Newman's "Networks An Introduction" [112] offers a good explanation of both of these centrality measures.

Subgraph Centrality is the measure based upon the adjacency matrix which we shall make use of in later chapters. It was introduced by Estrada and Rodriguez-Velazquez and counts the sum of the walks of each length which start and finish at the same node, where the weighting factor is the reciprocal of the factorial [48]. The following power series of the adjacency matrix, A, of a network converges to the corresponding matrix exponential,

$$\sum_{l=0}^{\infty} \frac{A^l}{l!} = \exp(A). \tag{2.11}$$

Obviously,

$$\exp(A) = \sum_{l=0}^{\infty} \frac{A^{2l}}{(2l)!} + \sum_{l=0}^{\infty} \frac{A^{2l+1}}{(2l+1)!} = \cosh(A) + \sinh(A), \tag{2.12}$$

where the first term accounts for the weighted sum of even-length walks and the second one accounts for odd-length walks in the network.

Definition 2.18. The **subgraph centrality** of a node, i, is given by $(\exp(A))_{ii}$.

This definition captures the participation of a node in all subgraphs of the network, with more weight given to the smaller ones. The key benefit of the subgraph centrality is that it takes into account all subgraphs that a node participates in regardless of how tangentially, while still retaining the focus on the local environment. As such it has been used to study landscape connectivity [44] and transport networks [160].

In the Friends network in Figure 2.1 the subgraph centralities of nodes 1 and 5 are 11.72 while the other nodes give a result of 6.63. The network in Figure 2.2 is a more interesting case. Here the top 6 nodes are node 3 with a result of 12.92 followed by a four way tie between 11, 12, 13 and 14 at 11.4514 with node 2 having the 6^{th} highest score at 10.2731. This result is a very different to the one for the closeness centrality. The subgraph centrality and other related centralities which essentially count walks, like Katz centrality, eigenvector centrality and to a lesser extent degree centrality are more a measure of how well connected a node is within its immediate vicinity and the density of the connections in that neighbourhood than whether or not it is a convenient place to be or a choke point. This time imagine that the nodes in Figure 2.2 represent academic publications or internet web pages and the links between them represent whether or not a given web page cites another. Generally, a web page or publication is given more credence if it has been cited by many other pages or publications and beyond that if it has been cited by web pages that are themselves reputable i.e. have also been cited

many times. The subgraph centrality accounts for this notion of importance because a node which is in a dense cluster of other nodes which rely on each other is likely to be the start and end point of a large number of short walks.

Obviously, as with all network based rankings, a high centrality can be both good and bad. For instance, four of the five top ranked nodes from Figure 2.2 are only directly adjacent to one another and to the top node. This situation creates an impression of authority which could be warranted (NHS or other healthcare websites connecting to each other) but also may not be (conspiracy theory websites parroting each other).

The example given above is not perfect because in reality both academic citations and web pages are directed networks. A link from 1 web page to another does not imply that a reverse link exists and there is definitely no such return link in academic publications. However, it is still illustrative as an example.

We shall extend some of these centrality measures to the case of simplicial complexes in Section 3.4.

2.3 Laplacians: What Flows On A Network?

2.3.1 Connected Components and Network Laplacians

The Laplacian, L, of a network can be thought of as a network analogue of the Laplacian operator ∇^2 and describes flows on a network. In Chapters 6.13 and 6.14 of [112] Newman offers a very good introduction to the topic of the network Laplacian which also introduces its applications to random walks, resistor networks and diffusion which are beyond the scope of what we need them for.

Definition 2.19. The **Laplacian matrix** of a network is the matrix defined as follows [31, p. 2]:

Let $v_i, v_j \in V$ be two nodes in a network. Then, the Laplacian matrix has entries defined by

$$L_{ij} = \begin{cases} k_i & \text{if } i = j; \\ -1 & \text{if } i \neq j \text{ and there is an edge } \{i, j\}; \\ 0 & \text{otherwise.} \end{cases}$$

The Laplacian matrix can also be calculated from the adjacency matrix, A, of a network if we let D be a diagonal matrix whose entries are given by the degrees of the nodes. That is

$$D = \begin{pmatrix} k_1 & 0 & 0 & \dots \\ 0 & k_2 & 0 & \dots \\ 0 & 0 & k_3 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.13}$$

Then we can calculate the network Laplacian as

$$L = D - A. \tag{2.14}$$

The eigenvalues of a network's Laplacian matrix have an interesting connection to the number of connected components of said network. Take any row, i, of the Laplacian matrix and sum the entries

$$\sum_j L_{ij} = \sum_j D_{ij} - \sum_j A_{ij} = k_i - k_i = 0. \tag{2.15}$$

This evaluation is done by decomposing the Laplacian into its constituent parts. There is only one non-zero entry on each row of D so clearly $\sum_j D_{ij} = k_i$ and $\sum_j A_{ij} = k_i$ as before. From here we have $L\mathbf{1} = 0\mathbf{1}$ which means that $\mathbf{1}$ is an eigenvector of L with eigenvalue 0 for every network as per Definition 2.10, where $\mathbf{1}$ is the vector of all 1s of length n and n is the number of nodes in the network. This result can be expanded upon.

Theorem 2.20. *Let $G = (V, E)$ be a network with x connected components. Then the 0-eigenvalue of the Laplacian of that network has multiplicity x [100, 112].*

2.3.2 The Edge Incidence Matrix

There is another way of constructing the Laplacian matrix of a network through something called the edge incidence matrix [112, p. 155].

Definition 2.21. Let G be a network with n vertices and m edges, for each edge we arbitrarily define that the node with the lower index is end 1 and the node with the higher index is end 2. Then the **edge incidence matrix**, B, is an $n \times m$ matrix with entries given by

$$B_{ij} = \begin{cases} -1 & \text{if end 1 of edge } j \text{ is attached to vertex } i \\ 1 & \text{if end 2 of edge } j \text{ is attached to vertex } i \\ 0 & \text{otherwise.} \end{cases}$$

For example the edge incidence matrix of the network from Figure 2.1 is given in Table 2.1.

	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{1,5\}$	$\{1,6\}$	$\{2,3\}$	$\{2,5\}$	$\{3,5\}$	$\{4,5\}$	$\{4,6\}$	$\{5,6\}$
1	-1	-1	-1	-1	-1	0	0	0	0	0	0
2	1	0	0	0	0	-1	-1	0	0	0	0
3	0	1	0	0	0	1	0	-1	0	0	0
4	0	0	1	0	0	0	0	0	-1	-1	0
5	0	0	0	1	0	0	1	1	1	0	-1
6	0	0	0	0	1	0	0	0	0	1	1

Table 2.1: We can see here that each column of the matrix has exactly one 1 and one −1.

Theorem 2.22. *The Laplacian matrix, L, of a network can be calculated as $L = BB^T$ [112, p. 155].*

In Section 2.6, on Topology, we will use higher dimensional equivalents of the network Laplacian to calculate the topology of simplicial complexes in the same way that Theorem 2.20 allows us to calculate the number of connected components of a network. These topological calculations will allow us to investigate gaps in coverage of Wireless Sensor Networks in Chapter 5.

2.4 Models of Disease

Quantities of a substance are not the only things which can flow around a network. Another thing that can flow around a network is disease. Newman's "Networks: An Introduction" [112] provides a very good introduction to this topic in Chapter 17. We assume knowledge of SIS (Susceptible Infected Susceptible) and SIR (Susceptible Infected Recovered) models of disease spread under the assumption of a fully mixed population which are covered in [5, 11, 112]. We introduce the network based models below. In Chapter 6 we will investigate the effect of the underlying geometry of a Random Rectangular Annular Network on whether or not a given disease will become an epidemic.

2.4.1 Susceptible Infected Susceptible

The assumption that a population is fully mixed, that any individual is equally likely to have contact with any other, simplifies the mathematics in the study of disease spread. We know this model is not realistic because people are much more likely to pass a disease on to their friends, family and work colleagues than they are to strangers that they meet in the street. Similarly, plants are more likely to pass diseases on to other plants that are close by than ones that are further away. To address this limitation we can introduce a network-theoretic SIS model.

Definition 2.23. In the **SIS model (Susceptible Infected Susceptible) on a network** we let $s_i(t)$ be the probability that a node, i, is susceptible at time t and $x_i(t)$ be the probability that said node is infected at time t [112, p. 669]. We also assume that each node is constantly in contact with its neighbouring nodes. Each individual can only catch the disease from a neighbour, j, and so the probability of becoming infected in a given time between t and $t + dt$ is given by the probability that node i is infected, s_i, multiplied by the probability, β, of a contact resulting in an infection multiplied by the sum of the probabilities that the neighbouring nodes are infected, $\sum_j A_{ij}x_j$, where A is the adjacency matrix of the network. We assume that a node that was infected returns to the susceptible population at a rate, γ, so the change

in the probability that node, i, is susceptible because it has recovered is given by γx_i. The following pair of equations governing the system:

$$\frac{ds_i}{dt} = -\beta s_i \sum_j A_{ij} x_j + \gamma x_i, \tag{2.16}$$

$$\frac{dx_i}{dt} = \beta s_i \sum_j A_{ij} x_j - \gamma x_i, \tag{2.17}$$

where $s_i + x_i = 1$.

A key question in the study of disease is whether or not a disease will spread beyond the first few individuals who contract it. That is, whether or not it will become an epidemic. In a fully mixed model if the average person passes on the disease to more than one other person before recovering themselves then the disease will spread and if they pass it on to fewer than one person then the disease will die out.

For a given disease spreading on a network then whether or not it becomes an epidemic depends not only on the probability of infection and the probability of recovery but also on the structure of the network.

Theorem 2.24. *The epidemic threshold for a network with SIS disease dynamics is given by* $\frac{\beta}{\gamma} = \frac{1}{\lambda_1}$ *where* λ_1 *is the largest eigenvalue of the adjacency matrix [112, p. 661 - 670].*

This theorem tells us that if we fix a particular disease with a probability of infection β_1 and probability of recovery γ_1. Then we take two networks one where the largest eigenvalue is λ_{1a} and another where it is λ_{1b}. If $\frac{1}{\lambda_{1a}} < \frac{\beta_1}{\gamma_1} < \frac{1}{\lambda_{1b}}$ then the disease will become an epidemic in the first network but not the second.

Gómez et al. [67] extended this result to networks with weighted edges and also the case where each node has a limited number of contacts they can make with their neighbours as opposed to the situation where it is assumed that at any given time-step each node is in contact with all of its neighbours. A criticism of using network

models to track the spread of disease is that humans are mobile and the contact that a person makes in one hour or day are likely to be different from the contacts made in the next. The Gomez et al. paper addresses this criticism to some extent but Sanatkar et al. [137] showed that this value for the epidemic threshold also holds in dynamic switching networks which are systems where the set of connections is allowed to swap randomly between different adjacency matrices. For the applications we analyse the nodes are static and so we do not require this level of detail, nonetheless it is interesting to know it exists.

2.4.2 Susceptible Infected Recovered

There are some diseases, for example chickenpox, against which a person or thing become immune, and therefore unable to be reinfected, once they have had it. The previously described SIS dynamics are inappropriate to describe such diseases and so we need to introduce a third "recovered" disease state to the model.

Definition 2.25. In the **SIR model (Susceptible Infected Recovered) on a network** we let $s_i(t)$ be the probability that a node, i, is susceptible at time t, $x_i(t)$ be the probability that said node is infected at time t and $r_i(t)$ be the probability that i is recovered at time t [112, p. 661 - 662]. We also assume that each node is constantly in contact with its neighbouring nodes. Each individual can only catch the disease from a neighbour, j, and so the probability of becoming infected in a given time between t and $t+dt$ is given by the probability that node i is infected, s_i, multiplied by the probability, β, of a contact resulting in an infection multiplied by the sum of the probabilities that the neighbouring nodes are infected, $\sum_j A_{ij}x_j$, where A is the adjacency matrix of the network. As before we assume that a node that was infected recovers at a rate, γ, so the change in the probability that node, i, is able to pass on the disease because it has recovered is given by γx_i. The following trio of equations govern the system:

$$\frac{ds_i}{dt} = -\beta s_i \sum_j A_{ij}x_j, \tag{2.18}$$

$$\frac{dx_i}{dt} = \beta s_i \sum_j A_{ij} x_j - \gamma x_i, \tag{2.19}$$

$$\frac{dr_i}{dt} = \gamma x_i, \tag{2.20}$$

where $s_i + x_i + r_i = 1$.

It is possible to demonstrate in the same way as for SIS models that the epidemic threshold is also given by $\frac{1}{\lambda_1}$ where λ_1 is the largest eigenvalue of the adjacency matrix [112, Chapter 17.11].

There are other models of disease spread such as the SI model where individuals can only pass from susceptible to infected and never back again. Clearly in this case an epidemic is inevitable because the number of healthy or susceptible people can never decrease. The SIRS model describes a situation where the immunity conferred by recovering from the disease is not permanent and nodes lose their immune status at a fixed rate. Although these models are interesting we do not use them in later chapters and so do not consider them in more detail here.

2.5 Simplicial Complexes

2.5.1 Beyond Networks: What If The Situation Is A Little More Complicated?

Simplicial complexes have been much studied in the literature [71, 147] and definitions similar to those which appear in this section can be found elsewhere [66,97,107,108,153]. However, we repeat them here to make this thesis self-contained.

Definition 2.26. Let V be a set of nodes or vertices. Then a **k-simplex** is a set $\{v_0, v_1, \ldots, v_k\}$ such that $v_i \in V$ and $v_i \neq v_j$ for all $i \neq j$.

A **face** of a k-simplex is a $(k-1)$-simplex of the form $\{v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k\}$ for $0 \leq i \leq k$.

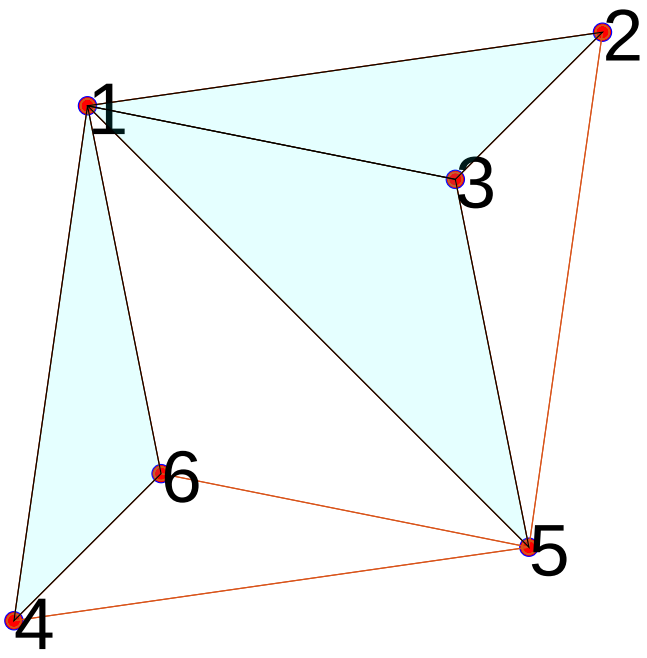


Figure 2.4: A simplicial complex representation of the Friends network.

A **simplicial complex** C is a collection of simplices such that if a simplex S is a member of C then all faces of S are also members of C. Less formally, a simplicial complex is a collection of simplices such that if $\{v_0, v_1, \ldots, v_k\}$ is a simplex then all of its faces $\{v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k\}$ are also simplices, and all of the faces of its faces $\{v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k\}$ are too, and so on down to the 0-simplices, which are formed by the nodes [62, p. 26].

As mentioned previously, networks satisfy the definition of simplicial complexes. The nodes are the 0-simplices which are specified by the set V, the edges are the 1-simplices and there are no higher order simplices. It is also possible to create simplicial complexes from networks.

Definition 2.27. A **clique complex** is a simplicial complex formed from a network as follows. The nodes of the network become the nodes of the simplicial complex. Let X be a clique of k nodes in the network. Then, X is a $(k-1)$-simplex in the clique complex [62, p. 30].

As an example in Figure 2.5 we illustrate a simplicial complex which has one

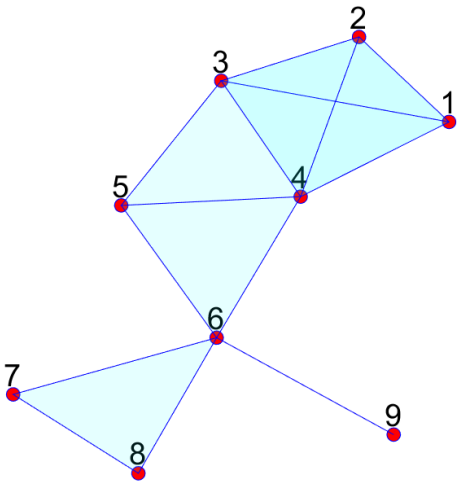


Figure 2.5: A clique complex with labelled nodes. This figure previously appeared in [50].

3-simplex $\{1,2,3,4\}$, seven 2-simplices $\{1,2,3\}$, $\{1,2,4\}$, $\{1,3,4\}$, $\{2,3,4\}$, $\{3,4,5\}$, $\{4,5,6\}$ and $\{6,7,8\}$. It also has fourteen 1-simplices represented by the edges and nine 0-simplices which are usually known as the nodes. We can see that this simplicial complex is a clique complex because every set of nodes which form a complete subgraph from the network are simplices in the simplicial complex.

However, the simplicial complex in Figure 2.4 is not a clique complex. We can see that the nodes $1,5,6$ form a triangle which is a clique on three nodes but they do not form a simplex within the simplicial complex. The lack of a simplex tells us that Ross lived with Rachel at some point in the series and at a separate point lived with Joey and at a third point Joey and Rachel lived together. Whereas, Monica, Chandler and Phoebe form a 2-simplex which tells us that they all lived in the same flat at the same time at least once during the series. We can see that the simplicial complex representation gives us more information on the, admittedly simple, system we are modelling than the equivalent network representation.

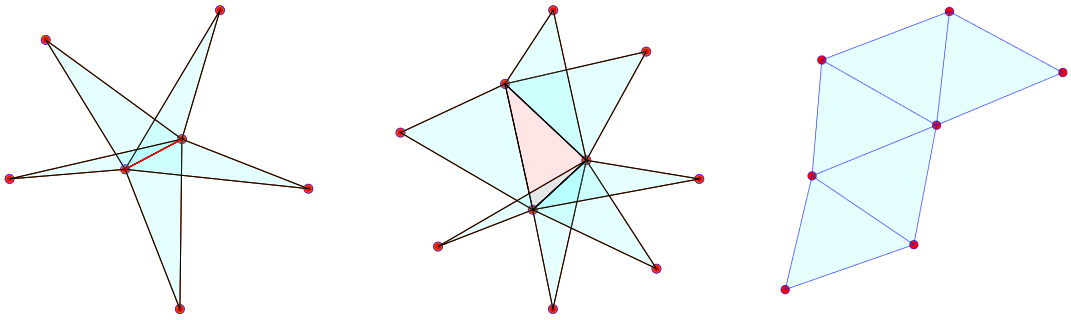
2.5.2 Face Facts: When Are Two Simplices Adjacent?

In network theory it is fairly clear when two nodes are adjacent. However, adjacency is less easy to define in simplicial complexes. There are multiple ways in which two simplices in a simplicial complex could be said to be adjacent. The first to be considered was Q–analysis which was introduced by Atkin [9] in 1972. Two simplices of any size are q–adjacent if they share a common face of dimension q. This notion of adjacency was originally introduced in terms of q–connectivity where two simplices of any size are q–connected if there is a series of simplices between them such that each simplex in the series is q–adjacent to the simplices that follow and precede it. We are focusing on adjacency between simplices of the same size and present below the concepts of upper and lower adjacency which were introduced by Goldberg in 2002 [66] although lower adjacency is a special case of Atkin's q–adjacency.

Definition 2.28. Let σ_j and σ_i be two k-simplices. They are **lower adjacent** if they share a common face. That is, for two distinct k-simplices $\sigma_j = \{v_0, v_1, \ldots, v_k\}$ and $\sigma_i = \{w_0, w_1, \ldots, w_k\}$ then σ_j and σ_i are lower adjacent if and only if there is a $(k-1)$-simplex $\tau = \{x_0, x_1, \ldots, x_{k-1}\}$ such that $\tau \subset \sigma_j$ and $\tau \subset \sigma_i$. We denote lower adjacency by $\sigma_j \backsim \sigma_i$.

For instance, in the simplicial complex in Figure 2.5, the 1-simplices $\{6,7\}$ and $\{6,9\}$ are lower adjacent because the 0-simplex $\{6\}$ is a common face and we can write $\{6,7\} \backsim \{6,9\}$. Similarly, $\{1,3,4\} \backsim \{3,4,5\}$ because they share the common face $\{3,4\}$. However, $\{4,5,6\}$ and $\{6,7,8\}$ are not lower adjacent because although they have the common 0-simplex $\{6\}$ they would need to share a common 1-simplex to be lower adjacent. Note that two 0-simplices can never be lower adjacent as we do not allow $\varnothing$ to be a -1-simplex.

Definition 2.29. Let σ_j and σ_i be two k-simplices. They are **upper adjacent** if they are both faces of the same common $(k+1)$-simplex. That is, for $\sigma_j = \{v_0, v_1, \ldots, v_k\}$ and $\sigma_i = \{w_0, w_1, \ldots, w_k\}$ then σ_j and σ_i are upper adjacent if and only if there is a $(k+1)$-simplex $\tau = \{x_0, x_1, \ldots, x_{k+1}\}$ such that $\sigma_j \subset \tau$ and $\sigma_i \subset \tau$. We denote the upper adjacency by $\sigma_j \frown \sigma_i$.



(a) (b) (c)

Figure 2.6: Illustration of the simplicial complexes S_5^2 (a), $t^2(1,2,4)$ (b) and P_5^2 (c). See text for definitions and notation. This figure previously appeared in [50].

In the simplicial complex in Figure 2.5, the 1-simplices $\{5,6\}$ and $\{4,6\}$ are upper adjacent because they are both faces of the 2-simplex $\{4,5,6\}$. So we can write $\{5,6\} \frown \{4,6\}$. Similarly, $\{1,3,4\} \frown \{2,3,4\}$ are upper adjacent as they are both faces of the 3-simplex $\{1,2,3,4\}$. However, $\{4,5,6\}$ is not upper adjacent to any other simplex as it is not part of any 3-simplices.

Also note that $\{6\} \frown \{7\}$ are upper adjacent because they are both faces of $\{6,7\}$. So two 0-simplices are upper adjacent if they are both faces of a 1-simplex which is identical to saying that two nodes are adjacent if they are connected by an edge in the network-theoretic sense. Hence upper adjacency of 0-simplices is the same as network-theoretic adjacency.

Goldberg also related these two concepts.

Theorem 2.30. *If two distinct k-simplices σ_i and σ_j are upper adjacent then they are also lower adjacent [66].*

2.5.3 Families of Simplicial Complexes

We shall now introduce some families of simplicial complexes which have been defined previously by Horak and Jost [70] which shall be important later. Firstly, we introduce the family denoted S_l^k. The simplicial complex S_l^k consists of a central $(k-1)$-simplex which is a face of every one of the l k-simplices. There are no other simplices except

those necessary by the closure axiom. For instance, S_l^2 would consist of an edge $\{1,2\}$ and l triangles of the form $\{1,2,i\}$ in addition to all subsimplices necessary by the closure axiom and S_l^1 consists of a central node with l pendant nodes connected to it, which corresponds to the star graph in graph theory. The simplicial complex S_5^2 is shown in Figure 2.6(a) with the central 1-simplex in red.

Next we introduce a family of simplicial complexes labelled $t^k(x_1, x_2, \ldots, x_{k+1})$ which consists of a central k-simplex with x_1 k-simplices lower adjacent through one face, x_2 k-simplices lower adjacent through another, and so on. A k-simplex which is lower adjacent to the central k-simplex can only be lower adjacent to other k-simplices which are lower adjacent to the central k-simplex through the same face as itself. There are no other simplices except those necessary by the closure axiom. One member of this family of simplices, $t^2(1,2,4)$ is shown in Figure 2.6(b) with the central 2-simplex in red.

The final family of simplicial complexes which we shall introduce are denoted P_l^k, consisting of a k-simplex at one end which is only adjacent to one other k-simplex which is only lower adjacent to the first k-simplex and another k-simplex, and so on until arriving at another end k-simplex. In addition, there are l k-simplices in the simplicial complex and no other simplices except those necessary by the closure axiom. Note that a simplicial complex P_l^1 is the same as a path graph in the traditional network theory. The simplicial complex P_5^2 is illustrated in Figure 2.6(c).

2.6 Topology

When we consider the simplicial complexes introduced in the previous section it is natural to think of the 2-simplices as being surfaces and 3-simplices being three dimensional objects taking up space. When we form simplicial complexes we could for instance stick 2-simplices together into the form of a ball with a void, or empty three dimensional space, in the middle. Similarly, we can arrange 1-simplices into a ring without 2-simplices in the centre so that if it was an actual object we could pass a different object through the centre of this ring or hole. Think of moving a ball bearing through the centre of a doughring or a coffee mug which are the canonical examples

	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

Table 2.2: The effect of multiplication in the Klein-four group.

of this phenomenon. In this section we present a method for describing the number of unique holes or voids there are for a given simplicial complex and an extension of the concept of the network Laplacian to higher levels of a simplicial complex which will allow us to easily calculate this number.

2.6.1 Group Theory

Before we can describe the topological space formed by a simplicial complex we need a few background details on Group Theory. We assume undergraduate level knowledge of Group Theory including the definitions of groups, subgroups, Abelian groups, cosets of the subgroups of a group and the fact that these cosets partition the group. Norman Biggs' book Discrete Mathematics [18] covers these topics.

For more detail on the advanced material which will allow us to define the Chain Complexes in the next section then Roman's Fundamentals of Group Theory: An Advanced Approach [131] could be consulted. Both of these books were a source for the examples, theorems and definitions detailed below.

Some examples of groups include the integers under addition, $\mathbb{Z}$ and the Klein-four group, $K_4 = \{e, a, b, c\}$ which is defined by the multiplication in Table 2.2 [18].

Subgroups of both of these groups exist. The subset $\{e, a\}$ is a subgroup of K_4 because $aa = e = ee \in \{e, a\}$ so we have the inverses property and $ae = ea = a \in \{e, a\}$ which gives us the closure property.

Additionally, the integers which are multiples of 3, $3\mathbb{Z} = \{3p; p \in \mathbb{Z}\}$ is a subgroup of $\mathbb{Z}$ under addition. Let $p, q \in \mathbb{Z}$ then $3p + 3q = 3(p + q) \in 3\mathbb{Z}$ and $3p + 3(-p) = 3(-p) + 3p = 0$ so $3\mathbb{Z}$ is closed under inverses and addition [18]. However, the odd numbers, $R = \{2p+1; p \in \mathbb{Z}\}$ are not a subgroup because it is not closed under addition. For $p, q \in \mathbb{Z}$ we have $2p + 1 + 2q + 1 = 2p + 2q + 2 = 2(p + q + 1) \notin R$.

Definition 2.31. Let $H \leq G$ be a subgroup of G. Then we denote the set of left cosets of H in G by G/H [131, p. 61].

The cosets of $\{e, a\} \subset K_4$ are $\{e, a\}$ and $\{b, c\}$ because $be = b$, $ba = c$, $ce = c$ and $ca = b$. The cosets of $3\mathbb{Z} \subset \mathbb{Z}$ are $3\mathbb{Z}$ itself, $(1+3\mathbb{Z}) = \{1+3p; p \in \mathbb{Z}\}$ and $(2+3\mathbb{Z}) = \{2+3p; p \in \mathbb{Z}\}$.

From here we can define a binary operation on the cosets of a subgroup H.

Definition 2.32. Let $H \leq G$ be a subgroup of G with the binary operation, $\circledast$, and let $a_1, a_2 \in G$ then the **coset product** of the cosets a_1H and a_2H is given by

$$a_1H \circledast a_2H = (a_1 \circledast a_2)H. \tag{2.21}$$

[131, p. 65]

Definition 2.33. Let G be a group and $N \leq G$ be a subgroup of G. Then N is a **normal subgroup** of G if for all $a \in G$ we have

$$aN = Na. \tag{2.22}$$

[131, p. 65]

Theorem 2.34. *Let $H \leq G$ be a subgroup of G under $\circledast$. Then G/H is a group under the coset product if and only if H is a normal subgroup of G.*

*In this situation G/H is known as the **quotient group** [131, p. 66 - 67].*

Both K_4 and $\mathbb{Z}$ are Abelian groups which means that all of their subgroups are normal by definition which means that $K_4/\{e, a\}$ and $\mathbb{Z}/3\mathbb{Z}$ are both quotient groups.

It is possible to derive all the rest of the members of a group just through combining certain elements via the binary operation.

Definition 2.35. Let G be a group and $X \subseteq G$. Then

$$\langle X \rangle = \{a_1^{i_1} \dots a_b^{i_b} : a_1, \dots, a_b \in X \text{ and } i_1, \dots, i_b \in \mathbb{Z}\}. \tag{2.23}$$

That is $\langle X \rangle$ consists of all of the possible combinations of the elements of X under the binary operation of G and the use of inverses. If it is the case that $\langle X \rangle = G$ then we say that X is a **generating set** for G [131, p. 34 - 35].

The **rank** of a group, G is the size of the smallest possible generating set [79] for that group

$$Rank(G) = \min\left(\|X\| : \langle X\rangle = G\right). \tag{2.24}$$

We can see that $\{1\}$ is a generating set for $\mathbb{Z}$ and so $Rank(\mathbb{Z}) = 1$. For K_4 a generating set is given by $\{a, b\}$ because we can get $c = ab$ however if we were to just multiply copies of a several times we would alternate between a and e and so we cannot generate the rest of K_4 just from $\{a\}$. The same applies to $\{b\}$ and $\{c\}$ and so $Rank(K_4) = 2$.

2.6.2 Chain Complexes

We now introduce a group based on the simplices of a simplicial complex [4,39,62,107].

Definition 2.36. We can create a k**-chain** by summing the k-simplices, σ_i, of a simplicial complex [39, p66] as per the formula

$$c = \sum_{i=1}^{n} b_i \sigma_i \tag{2.25}$$

where $b_i \in F$ are values from a field F and n is the number of simplices in the simplicial complex.

For all examples we use the field $\mathbb{Z}$ for the sake of simplicity which is all that is required for the applications explored here but it is possible to use other fields such as $\mathbb{Z}/2\mathbb{Z}$ or $\mathbb{R}$ [66]. All definitions in this section apply generally regardless of which field is used.

Definition 2.37. For a simplicial complex, S, the k^{th} **chain group** is the collection of all possible k-chains of S

$$C_k = \left\{ \sum_{i=1}^{n} b_i \sigma_i : \sigma_i \in S, b_i \in F \right\}. \tag{2.26}$$

[39, p. 66]

We can see that the identity of a chain group is $\sum_{i=1}^{n} 0\sigma_i = 0$, we inherit associativity from addition in the field and closure comes from it not being possible to build extra simplices in the simplicial complex. We also have that $(\sum_{i=1}^{n} b_i\sigma_i)^{-1} = -\sum_{i=1}^{n} b_i\sigma_i$ and inherit commutativity from the field which makes the chain group Abelian.

Definition 2.38. Let $\{v_0, \ldots, v_k\}$ be a simplex in a simplicial complex. An **orientation** of this simplex is any ordering of its vertices $[v_0, \ldots, v_k]$. We additionally define that two orientations of a simplex are the same if they differ by an even number of swaps of the positions of its elements and they are inverses of each other if they differ by an odd number of position swaps [62, p. 66].

We can demonstrate this swaps rule with this example

$$\begin{aligned} [v_0, v_1, v_2, v_3, \ldots, v_k] &= -[v_2, v_1, v_0, v_3, \ldots, v_k] \\ [v_0, v_1, v_2, v_3, \ldots, v_k] &= [v_2, v_0, v_1, v_3, \ldots, v_k]. \end{aligned} \tag{2.27}$$

In the first case there has been one swap of v_2 and v_0 while in the second case there were two swaps, with the additional swap being v_1 and v_0. We primarily work with simplicial complexes where we have numbered nodes and so we define the orientation on our simplices to be the natural one where the nodes are arranged in ascending order.

We can see that there is a chain group at each level of the simplicial complex. We can pass from groups at one level to the next through the use of the boundary operators and the whole set of groups and these functions is known as a chain complex.

Definition 2.39. The **boundary operator** is a function, $\partial : C_k \to C_{k-1}$, which acts on an oriented k-simplex, $\sigma_k = [v_0, \ldots, v_{i-1}, v_i, v_{i+1}, \ldots, v_k]$ as

$$\partial_k \sigma_k = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k]. \tag{2.28}$$

A k-chain which is the result of applying the boundary operator to a $(k+1)$-chain is called a **boundary** [39, p. 67].

What the boundary operator does is map a simplex to the sum of its faces with the orientation of each face defined by which node is being removed in relation to the

position of that node in the orientation of the original simplex. When a node in an odd position is removed the orientation of the face is negative and when a node in an even position is removed then it is positive.

This fact gives the boundary operator a nice property which is vital to using these groups to understand the topology of a simplicial complex.

Theorem 2.40. *Let $\sigma_k = [v_0, \ldots, v_k]$ be an oriented k-simplex then*

$$\partial_{k-1}\partial_k\sigma_k = 0. \tag{2.29}$$

[39, p. 67]

The boundary of any $(k+1)$-simplex is an example of a k-cycle.

Definition 2.41. A **k-cycle** is any k-chain which has boundary 0 [39, p. 67].

Let S be a simplicial complex with n k-simplices. Then the set of all k-cycles is denoted by Z_k,

$$Z_k = ker(\partial_k) = \left\{ \sum_{i=1}^{n} b_i\sigma_i \in C_k : \partial_k \left(\sum_{i=1}^{n} b_i\sigma_i \right) = 0 \right\}. \tag{2.30}$$

We can see that Z_k is closed under inverses and addition within C_k and so $Z_k \leq C_k$. Consider the simplicial complex in Figure 2.4, we have that $[1,2] - [1,3] + [2,3] + [1,4] - [1,6] + [4,6] \in Z_1$ as demonstrated in Equation (2.31). This 1-cycle is the result of $\partial_2([1,2,3] + [1,4,6])$.

$$\begin{aligned} &\partial_1([1,2] - [1,3] + [2,3] + [1,4] - [1,6] + [4,6]) \\ =&[2] - [1] - [3] + [1] + [3] - [2] + [4] - [1] - [6] + [1] + [6] - [4] \\ =&0 \end{aligned} \tag{2.31}$$

The boundaries of $(k+1)$-chains are not the only members of Z_k. In Equation (2.32) we demonstrate that $[2,3] - [2,5] + [3,5] \in Z_1$ despite not being the boundary of any 2-chain.

$$\partial_1([2,3]-[2,5]+[3,5]) = [3]-[2]-[5]+[2]+[5]-[3] = 0 \tag{2.32}$$

Finally in Equation (2.33) we can see that $[1,2]+[1,3]+[2,3]+[1,4] \notin Z_1$.

$$\begin{aligned} &\partial_1([1,2]+[1,3]+[2,3]+[1,4]) \\ =&[2]-[1]+[3]-[1]+[3]-[2]+[4]-[1] \\ =&2[3]+[4]-3[1] \neq 0 \end{aligned} \tag{2.33}$$

We can see that 1-cycles consist of sums of 1-simplices which contain each node an equal number of times with opposing orientations. It is true more generally that k-cycles consist of sums of k-simplices which contain each $(k-1)$-face an equal number of times with opposing orientations.

We can define another subgroup of C_k which consists of just the boundaries of the $(k+1)$-chains.

Definition 2.42. The **image** of ∂_{k+1} is the set of all k-cycles that are boundaries of $(k+1)$-chains [39, p. 68]. Let σ_i denote a k-simplex of a simplicial complex, S, and let τ_j denote a $(k+1)$-simplex of S. We denote this set by

$$B_k = img(\partial_{k+1}) = \left\{ \sum_{i=1}^{n} a_i\sigma_i \in C_k : \partial_{k+1}\left(\sum_{j=1}^{m} b_j\tau_j\right) = \sum_{i=1}^{n} a_i\sigma_i \text{ where } \sum_{j=1}^{m} b_j\tau_j \in C_{k+1} \right\}. \tag{2.34}$$

Due to the fact that all elements in B_k are also elements of Z_k as demonstrated by Theorem 2.40 and B_k is closed under inverses and addition of simplices then $B_k \leq Z_k$. It is also normal due to the binary operation in C_k being commutative. By definition we can generate all of the members of B_1 by adding together the boundaries of the 2-simplices. Generating all of the members of B_1 for the simplicial complex in Figure 2.4 shows that any member of B_1 is a result of summing $[1,2]-[1,3]+[2,3]$, $[1,3]-[1,5]+[3,5]$ and $[1,4]-[1,6]+[4,6]$. We can use this result to introduce a new group which will allow us to discuss the topology of the simplicial complexes.

Definition 2.43. The quotient group $H_k = Z_k/B_k$ is called the k^{th} **homology group** of the simplicial complex [39, p. 69].

The identity of the k^{th} homology group is the coset containing the members of B_k as is natural from Theorem 2.34 which means that the result of adding a member of B_k to a representative of any coset in H_k is a representative of the same coset.

Recall that we are trying to find a way to describe the number of holes or voids of a simplicial complex. In the simplicial complex in Figure 2.4 we can start by considering that there are five 1-cycles in Z_1 which consist of three 1-simplices and do not form the boundary of any 2 simplex. These 1-cycles are $[1,2]-[1,5]+[2,5]$, $[2,3]-[2,5]+[3,5]$, $[1,5]-[1,6]+[5,6]$, $[1,4]-[1,5]+[4,5]$ and $[4,5]-[4,6]+[5,6]$. But it would not be natural to say that there are five holes in this simplicial complex. Firstly, we can show that $[1,2]-[1,5]+[2,5]$ and $[2,3]-[2,5]+[3,5]$ are part of the same coset and therefore represent the same hole. We can add elements of B_k to $[1,2]-[1,5]+[2,5]$ to transform it into a scalar multiple of $[2,3]-[2,5]+[3,5]$

$$\begin{aligned}
&([1,2]-[1,5]+[2,5])-([1,2]-[1,3]+[2,3])-([1,3]-[1,5]+[3,5])\\
=&(-[1,5]+[2,5]+[1,3]-[2,3])-([1,3]-[1,5]+[3,5])\\
=&-([2,3]-[2,5]+[3,5]).
\end{aligned} \tag{2.35}$$

The other issue with using the aforementioned definition for the number of holes is that it would ignore any holes which were bounded by more than three 1-simplices. Consider for example a new simplicial complex which consists of four 1 simplices, $[a,b]$, $[a,d]$, $[b,c]$ and $[c,d]$ and the 0-simplices from the closure axiom. There would clearly be a hole in the simplicial complex but it would not be counted if we only considered 1-cycles in Z_1 which consist of three 1-simplices and do not form the boundary of any 2 simplex.

Below we categorise the members of H_1 for the simplicial complex in Figure 2.4 where $x, y, z \in \mathbb{Z}$.

$$
\begin{aligned}
&B_1\\
&x([1,5]-[1,6]+[5,6])+B_1\\
&x([2,3]-[2,5]+[3,5])+B_1\\
&x([4,5]-[4,6]+[5,6])+B_1\\
&x([1,5]-[1,6]+[5,6])+y([2,3]-[2,5]+[3,5])+B_1\\
&x([1,5]-[1,6]+[5,6])+y([4,5]-[4,6]+[5,6])+B_1\\
&x([2,3]-[2,5]+[3,5])+y([4,5]-[4,6]+[5,6])+B_1\\
&x([1,5]-[1,6]+[5,6])+y([2,3]-[2,5]+[3,5])+z([4,5]-[4,6]+[5,6])+B_1
\end{aligned}
\tag{2.36}
$$

It would also not be sensible to claim that the number of cosets in H_1 is the number of holes in the simplicial complex. Firstly, there are infinitely many cosets in H_1 and there are clearly not infinitely many holes in the simplicial complex in Figure 2.4. Furthermore, B_1 consists only of boundaries of combinations of 2-simplices and so clearly cannot contain any holes. Similarly, some cosets consist of multiple holes added together and so cannot be considered unique holes in their own right. For example, the 1-cycle $[1,4]-[1,5]+[4,5]$ belongs to $-1([1,5]-[1,6]+[5,6])+1([4,5]-[4,6]+[5,6])+B_1$

$$
\begin{aligned}
&[1,4]-[1,5]+[4,5]+B_1\\
=&[1,4]-[1,5]+[4,5]+[1,4]-[1,6]+[4,6]+B_1\\
=&-[1,5]+[4,5]+[1,6]-[4,6]+B_1\\
=&-[1,5]+[1,6]-[5,6]+[4,5]-[4,6]+[5,6]+B_1\\
=&-([1,5]-[1,6]+[5,6])+([4,5]-[4,6]+[5,6])+B_1.
\end{aligned}
\tag{2.37}
$$

Note though that none of $[1,5]-[1,6]+[5,6]+B_1$, $[2,3]-[2,5]+[3,5]+B_1$ or $[4,5]-[4,6]+[5,6]+B_1$ can be created by combining the other two elements and that every other element of H_1 can be created by combining these elements which means that

$$\langle\{[1,5]-[1,6]+[5,6]+B_1, [2,3]-[2,5]+[3,5]+B_1, [4,5]-[4,6]+[5,6]+B_1\}\rangle = H_1. \quad (2.38)$$

Equation (2.38) gives $Rank(H_1) = 3$. When we consider the simplicial complex in Figure 2.4 we would expect it to have 3 holes. This number is also referred to as the k^{th} Betti number of the simplicial complex and can be calculated with a formula.

Definition 2.44. The k^{th} **Betti number**, β_k, of the simplicial complex is the count of unique $(k+1)$-dimensional holes, or

$$\beta_k = Rank(H_k) = Rank(Z_k) - Rank(B_k). \quad (2.39)$$

[39, p. 69]

We have come to this definition by consideration of an example but it is true more generally because each coset of H_k consists of a summation of the unique holes in the simplicial complex plus every combination of the elements of B_k. Then we can generate all of these cosets by applying the coset product to the cosets containing precisely one unique hole. For the 0-simplices the "holes" are the connected components of the network the simplicial complex is built upon and for 2-simplices the holes represent 3-dimensional voids in the interior of the simplicial complex.

In order to use Equation (2.39) we need to be able to calculate $Rank(B_k)$ and $Rank(Z_k)$. Unfortunately, calculating $Rank(B_k)$ is more complicated than counting the number of k-simplices in a simplicial complex. Consider a new simplicial complex with four 2-simplices, $[1,2,3], [1,2,4], [1,3,4]$ and $[2,3,4]$ we can calculate

$$\begin{aligned} &\partial_2([1,2,3]) - \partial_2([1,2,4]) + \partial_2([1,3,4]) \\ =&[2,3] - [1,3] + [1,2] - [2,4] + [1,4] - [1,2] + [3,4] - [1,4] + [1,3] \\ =&[2,3] - [2,4] + [3,4] \\ =&\partial_2([2,3,4]). \end{aligned} \quad (2.40)$$

So in this case $Rank(B_2) = 3$ but there are four 2-simplices.

	$[1,2,3]$	$[1,3,5]$	$[1,4,6]$
$[1,2]$	1	0	0
$[1,3]$	-1	1	0
$[1,4]$	0	0	1
$[1,5]$	0	-1	0
$[1,6]$	0	0	-1
$[2,3]$	1	0	0
$[2,5]$	0	0	0
$[3,5]$	0	1	0
$[4,5]$	0	0	0
$[4,6]$	0	0	1
$[5,6]$	0	0	0

Table 2.3: The matrix representation of ∂_2 for the simplicial complex in Figure 2.4.

Fortunately, we can address these issues of how to calculate these ranks by noticing that C_k is also a vector space where we define multiplication in the vector space to be distributive across linear combinations of simplices. This definition means we can write ∂_k as an $m \times n$ matrix where there are m $(k-1)$-simplices and n k-simplices.

Definition 2.45. Let S be a simplicial complex with n k-simplices and m $(k-1)$-simplices, we can represent the boundary operator, ∂_k, by an $m \times n$ matrix where each column represents a k-simplex and each row a $(k-1)$-simplex [107]. The entries of each column are non-zero where the $(k-1)$-simplex representing a row is the face of the simplex in a given column, with -1 where the coefficient of the boundary operator for that face is negative and 1 the entry where it is positive. The entries of ∂_k are given by

$$(\partial_k)_{ij} = \begin{cases} -1 & \text{if the coefficient of } \sigma_{(k-1),i} \text{ is negative for } \partial_k(\sigma_{k,j}) \\ 1 & \text{if the coefficient of } \sigma_{(k-1),i} \text{ is positive for } \partial_k(\sigma_{k,j}) \\ 0 & \text{otherwise.} \end{cases}$$

For the 1-simplices the boundary operator and the edge incidence matrix from Definition 2.21 coincide. So we have already seen the matrix representation of ∂_1 for the simplicial complex in Figure 2.4. The matrix representation of ∂_2 is shown in Table 2.3.

We can verify from the matrix forms that $\partial_1 \partial_2 = 0$. We can also see that because each column of ∂_2 has the only non-zero entry in at least 1 row we can conclude that

	[1,2]	[1,3]	[1,4]	[1,5]	[1,6]	[2,3]	[2,5]	[3,5]	[4,5]	[4,6]	[5,6]
[1]	1	0	0	0	0	-1	-1	0	0	0	0
[2]	0	1	0	0	0	1	0	-1	0	0	0
[3]	0	0	1	0	0	0	0	0	-1	-1	0
[4]	0	0	0	1	1	0	1	1	1	1	0
[5]	0	0	0	0	1	0	0	0	0	1	1
[6]	0	0	0	0	0	0	0	0	0	0	0

Table 2.4: The row reduced form of the matrix representation of ∂_1 for the simplicial complex in Figure 2.4. The raw form can be seen in Table 2.1.

they are linearly independent and so all three are part of the smallest generating set for B_1 and $Rank(B_1) = 3$. To calculate the size of the smallest generating set for Z_1 we can apply row operations to ∂_1 to work out the form of any vector which would be sent to 0 under ∂_1. The number of free variables is the minimum size of a generating set of Z_1. The result of doing this evaluation is displayed in Table 2.4.

From here we can derive that any element of Z_1 has the form

$$\begin{pmatrix} f+g \\ h-f \\ i+j \\ k-g-h-i \\ -(j+k) \\ f \\ g \\ h \\ i \\ j \\ k \end{pmatrix} \tag{2.41}$$

There are six free variables and so $Rank(Z_1) = 6$ which means that $Rank(H_1) = 3$ as expected. Additionally, setting all other variables to 0 allows us to recover the boundaries of the three 2-simplices with $f = 1, h = 1$ and $j = 1$ respectively.

This case was quite straightforward because it was clear that the columns of ∂_2 were linearly independent and it was also quite a small example. A more general way of calculating the rank of the homology group at a particular level is through the calculation of the Smith normal form of the two matrices [40].

2.6.3 Hodge Laplacians

In the same manner that we can generalise the edge incidence matrix to the boundary operator at each level of the simplices we can also generalise the Laplacians to higher orders through the use of the Hodge Laplacians.

Definition 2.46. The Hodge Laplacian for the k-simplices of a simplicial complex can be calculated as

$$L_k = \partial_{k+1}\partial_{k+1}^T + \partial_k^T\partial_k. \tag{2.42}$$

It has entries defined by

$$(L_k)_{ij} = \begin{cases} deg_u(\sigma_i) + k + 1 & \text{if } i = j \\ 1 & \text{if } i \neq j, \sigma_i \nsucc \sigma_j, \text{ and } \sigma_i \sim \sigma_j, \text{ where the coefficient of their common face is the same} \\ -1 & \text{if } i \neq j, \sigma_i \nsucc \sigma_j, \text{ and } \sigma_i \sim \sigma_j, \text{ where the coefficient of their common face is different} \\ 0 & \text{otherwise.} \end{cases}$$

where $deg_u(\sigma_i)$ is the number of $(k+1)$-simplices of which σ_i is a face.

There are no -1-simplices to map the nodes to through a boundary operator and so $\partial_k = 0$ which means we recover the graph theoretic Laplacian as $L_0 = \partial_1\partial_1^T$. Recall from Thm 2.20 that the multiplicity of the 0-eigenvalue of the graph Laplacian gives the number of connected components of a graph. Eckmann showed, in 1944, that the Hodge Laplacian, L_k, could be used to determine β_k of a simplicial complex [38] (Horak and Jost presented an English language version [70]).

Theorem 2.47. *Let S be a simplicial complex with kth Betti Number, $\beta_k = x$. Then the 0-eigenvalue of the Hodge Laplacian, L_k, of the k-simplices of that simplicial complex multiplicity x. If there is no 0-eigenvalue of L_k then the Betti number at that level of the simplicial complex is 0.*

The Betti numbers in Chapter 5 were calculated in this manner. We present below the edge level and triangle level Laplacians for the simplicial complex in Figure 2.4.

$$L_1 = \begin{pmatrix} 3 & 0 & 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 3 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 3 & 1 & 0 & 1 & 0 & 1 & 0 & -1 \\ 1 & 1 & 0 & 1 & 3 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 3 & 1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 1 & 2 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 3 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 & 0 & 0 & 1 & 1 & 2 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & -1 & -1 & 1 & 2 \end{pmatrix} \tag{2.43}$$

$$L_2 = \begin{pmatrix} 3 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{2.44}$$

There are three 0-eigenvalues of L_1, none for L_2 and one for L_0 as expected.

Further information on the Hodge Laplacian matrices can be found in [66, 97, 107, 108, 153].

2.7 Statistics

In this section we explain and describe some statistical tests which we use in later sections to demonstrate some results.

2.7.1 AIC and BIC

The Akaike information criterion (AIC) [88,152] and the Bayesian information criterion (BIC) [88] are two methods by which different models can be compared when they are trying to predict the same data. They are defined as follows:

$$AIC = 2k - 2\ln\left(\hat{L}\right), \tag{2.45}$$

$$BIC = k\ln(n) - 2\ln\left(\hat{L}\right), \tag{2.46}$$

where n is the number of data points, k is the number of parameters to be estimated and $\hat{L} = p\left(x|\hat{\theta}, M\right)$ is the maximized value of the likelihood function of the model M, where $\hat{\theta}$ are the parameter values that maximize the likelihood function and x are the data points. For a series of models trying to describe the same dataset, the model with the smallest value AIC or BIC gives the best fit for the data. We can see that these criteria reward a model for predicting more of the data but penalise it for having more variables than are needed to predict the dataset. This reduces the risk of developing an incredibly complex model which predicts the model really well just by having many variables.

To use AIC or BIC we fit the dataset to each of the models. Then we rank each model according to their AIC and/or BIC which gives us the model which best fits the data. However, it may be the case that the difference between the smallest and second smallest AICs is very small and so random variation within the datasets could give an advantage to one model over another. We can compare the AIC of the best ranking model with model i as follows:

$$\Delta AIC_i = \exp\left(\left(AIC_{min} - AIC_i\right)/2\right) \tag{2.47}$$

where AIC_{min} is the AIC for the top model in the ranking.

If $\Delta AIC_i < 0.01$ we consider that the first model in the ranking is significantly different from model i. If this assessment is true for all i then we accept that there is evidence that the top model in the ranking better predicts the data than any of the other models.

When we compare the BIC results of different models we follow the same process but instead of applying Equation (2.47) we use the Kass-Raftery criteria [80] as follows:

ΔBIC_i	Meaning
0-2	Not significant
2-6	Positive
6-10	Strong
>10	Very strong

If the difference in the BIC values is less than or equal to 2, this criterion is not able to distinguish between the two models. If, however, it is between 6 and 10 there is strong evidence to consider the model with the smallest BIC as the most significant.

2.7.2 Spearman's Rank Correlation Coefficient

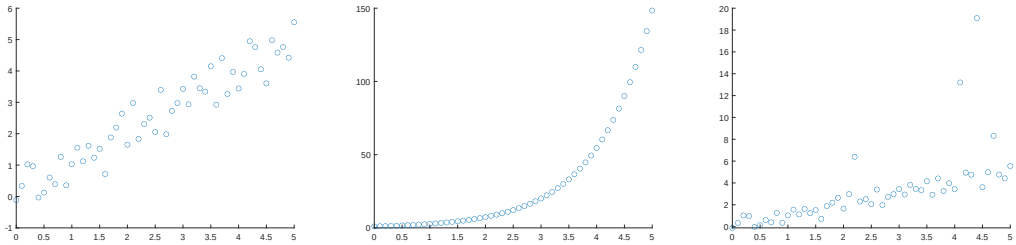
Spearman's rank correlation coefficient was introduced by Spearman in 1904 [149] based on the Pearson correlation coefficient which was introduced by Bravais in 1846 [21] and popularised by Pearson in 1896 [118].

Definition 2.48. Given paired random variables X, Y their **Pearson correlation coefficient** is given by

$$rp(X,Y) = \frac{cov(X,Y)}{sd_X sd_Y} \tag{2.48}$$

where $cov(X,Y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1}$ is the covariance between X and Y and $sd_X = \sqrt{\frac{\sum(x_i-\bar{x})^2}{N-1}}$ represents the standard deviation of X.

Take X and construct a new variable by replacing the value of each entry by its position in an ascending list of the values of X. Call this new variable $od(X)$. Do the same for Y. Then the **Spearman rank correlation coefficient** is given by



(a) (b) (c)

Figure 2.7: A linear relationship between two variables with experimental noise (a), An exponential relationship between two variables (b) and a relationship between two variables with large outliers (c).

$$rs(X,Y) = rp(od(X), od(Y)) = \frac{cov(od(X), od(Y))}{sd_{od(X)} sd_{od(Y)}}. \tag{2.49}$$

There are advantages and disadvantages to either choice of correlation coefficient. In the first instance the Pearson correlation simply gives more information about the relationship between two variables than Spearman's does as Spearman acknowledges himself [149]. Take Figure 2.7 for example, there is clearly a linear relationship between the two variables with some experimental noise in 2.7 (a). The Pearson coefficient of these two variables is 0.9507 and Spearman's gives 0.9539. However, in Figure 2.7 (b) the relationship between the two variables is exponential and Pearson's coefficient is 0.8544 while Spearman's version is 1. difference demonstrates that Spearman's coefficient is not sensitive to two variables which are related non-linearly whereas Pearson's is [149].

Furthermore, Spearman's rank correlation is more robust to large outliers than the traditional Pearson's correlation [149]. If we consider Figure 2.7 (c) we have a Spearman's result of 0.9244 and a Pearson's of 0.6586. We can conclude that where the magnitude of the differences between two variables is of importance then we would wish to use the Pearson's correlation coefficient but if we are just looking to assess whether or not there is a relationship (not necessarily linear) between two variables then the Spearman's correlation is preferable.

Compost grown plant yield (kg)	Straw grown plant yield (kg)
2.1	2.3
1.9	1.6
4.3	1.2
1.8	2.0
3.0	2.8
1.8	1.9
1.1	1.6
1.4	1.8
2.5	3.0
2.4	2.1

Table 2.5: The weight in kilos of the potato yield under different growing conditions.

2.7.3 Mann-Whitney U Test

The Mann-Whitney U test is a hypothesis test which is not dependent on the underlying distribution of the data, for instance in the way that the t-test requires that the underlying data is normally distributed [84]. For this reason it is useful to apply in situations where we have small sample sizes and may not necessarily be sure of the underlying distribution. It was introduced by Mann and Whitney in 1947 [99].

Imagine a situation where we wished to assess the difference in the yield between growing potatoes in compost compared to growing them in straw. We would grow perhaps 10 potato plants in compost and 10 in straw and measure the yield of each plant at the end of the season. We provide some example results in Table 2.5.

We demonstrate how to use the Mann-Whitney U test to establish whether the median yield of the two methods are different. We start by ordering the yields from largest to smallest. Table 2.6 contains the results of this process. Notice that we do not break ties but instead give each plant in a tie-break situation the average of the rankings they would occupy. For example the two plants with a yield of 3kg would be in positions 2 and 3 and so they are both given the ranking 2.5, we then carry on the ranking with 4 for the plant with a yield of 2.8kg.

The next step is to sum up the rankings for each sample the result of which is labelled R_i for sample i, we then calculate the U statistic for each sample via the formula

Compost grown plant yield ranks	Straw grown plant yield ranks
8.5	7
11.5	16.5
1	19
14	10
2.5	4
14	11.5
20	16.5
18	14
5	2.5
6	8.5

Table 2.6: The ranking of each plant in terms of the potato yield under different growing conditions.

	Compost grown plants	Straw grown plants
R	100.5	109.5
n	10	10
U	54.5	45.5

Table 2.7: The calculation of the U statistic for each of the samples.

$$U_i = n_1 n_2 + \frac{n_i(n_i+1)}{2} - R_i. \tag{2.50}$$

This process is detailed in Table 2.7. Note that the sums of the ranks for a sample of size n is $\frac{n(n+1)}{2}$ which is the lowest that R_i could be. The value of R_i would attain this minimum in the case that all data-points from one sample are less than all of the data points from the other sample and result in $U_i = n_1 n_2$. In this situation, the R_j result for the other sample would be given by $\frac{n_j(n_j+1)}{2}$ plus all n_j samples being above all n_i samples which gives $n_i n_j = n_1 n_2$, $R_j = \frac{n_j(n_j+1)}{2} + n_1 n_2$ and $U_j = 0$. In fact it is the case in general that $U_1 + U_2 = n_1 n_2$ and so we can conceive of the U statistic as a measure of how much the distribution of one sample is distinct from the distribution of the other.

We take $U = \min\{U_1, U_2\} = 45.5$ to be the overall U statistic. The U statistic is normally distributed with mean $\mu = \frac{n_1 n_2}{2}$ and standard deviation $\sigma = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$ (Chapter 6, [85]). We can translate our U statistic into a z-score

$$Z = \frac{U - \mu}{\sigma} = -\frac{4.5}{13.2288} = -0.3402. \tag{2.51}$$

This z-score can be turned into a p-value of 0.7337 and we can conclude that there is no evidence that the medians of the two samples are significantly different.

2.8 Computer Code

Where calculations have been done using simplicial complexes approaches, such as calculations of Betti numbers, the Javaplex [154] package was used which was developed by Tausz, Vejdemo-Johansson and Adams.

All code which has been used to make calculations in this work is available at `https://doi.org/10.15129/7d299611-6679-40f4-89e7-9a7c6e154e7e`.

Chapter 3

Concepts

This chapter has been published as part of [50]. The idea to extend centrality measures to the case of simplicial complexes was a contribution of Estrada to [50]. All of the definitions, proof and results in this chapter represent contributions by the author to [50]. Both Estrada and the author reviewed all aspects of [50].

3.1 Literature Review

The traditional network theory has ways of establishing centralities on higher dimensional entities in a model, for example the topological centrality [174] or the edge betweenness centrality [76], which can be used to narrow down which pieces of power grid infrastructure to monitor and mitigate against their failure. Unfortunately neither of these examples can be extended to cliques of three or more nodes interacting with other cliques with the same number of nodes.

It is possible to adapt Everett and Borgatti's [55] group centralities to cliques. They defined group degree, group closeness and group betweenness for a set of nodes. The group degree is calculated by taking a set of nodes in the network and calculating the number of other nodes in the network which are adjacent to at least one node in the set. For example, in the network in Figure 2.2 the set $\{1, 6\}$ has a group degree of 6 because the node 1 is adjacent to 4 other nodes $(2, 3, 10, 15)$ and node 6 is adjacent to 3 other nodes $(2, 5, 7)$ but node 2 is counted only once despite being adjacent to both. The group closeness is defined as the reciprocal of the sum of the minimum shortest

path distance from each node outside the set to a node in the set. Taking the set $\{1, 6\}$ from the network in Figure 2.2 again we know that the nodes $2, 3, 5, 7, 10$ and 15 have a minimum shortest path distance of 1 to $\{1, 6\}$. We can also see that nodes $8, 9, 11, 12, 13$ and 14 are at distance 2 from node 1 while nodes $4, 8, 9, 16$ and 17 are at distance 2 from node 6 which means that all of the nodes in the network other than six nodes which are directly adjacent to $\{1, 6\}$ have minimum shortest path distance to this set of 2. Therefore the group closeness of $\{1, 6\}$ is $\frac{1}{24}$. The group betweenness is derived from Freeman's graph theoretic betweenness [57] which counts, for a given node, a, the proportion of shortest paths between two other nodes that pass through a. The group betweenness extends this definition to the proportion of shortest paths between two nodes in the network which are not members of the set that pass through any node in the set. This extension is done in a way analogous to how group degree and group closeness extend the degree and closeness respectively. If we forced the groups to be complete subgraphs we would have viable centralities for triangles (a clique of three nodes) or tetrahedrons (a clique of four nodes). However, if we analyse a phenomenon as a simplicial complex then this definition would ignore the structure of the simplicial complex i.e. 2-simplices interacting with other 2-simplices.

There are a small number of centrality measures which are unique to the case of simplicial complexes which include the concepts of local homology by Robinson et al. [129] which estimates the effect on the simplicial complex of removal of a node and its neighbourhood.

Persistent Homology offers another way to define a centrality measure which is specific to the simplicial complex case through Homological Scaffolds which were used by Petri et al. [120] to investigate fMRI networks. Edelsbrunner offers a good introduction to Persistent Homology [39, Chapter 11] but a brief overview is provided here because it offers an alternative way of viewing the Random Geometric Simplicial Complexes that we will analyse in Chapter 5. Persistent Homology starts with a filtration of a simplicial complex. Let S be a simplicial complex, a filtration of S assigns each simplex $\sigma \in S$ a real number α and constructs a series of simplicial complexes S_γ where $\sigma \in S_\gamma$ if and only if $\gamma \geq \alpha$. Note that to satisfy the definition of a simplicial complex the

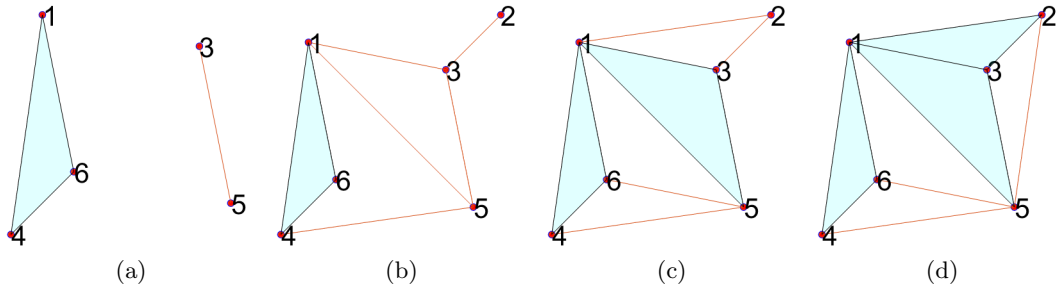
α	Simplex
0	$\{1\},\{3\},\{4\},\{5\},\{6\},\{1,4\},\{1,6\},\{4,6\},\{3,5\},\{1,4,6\}$
1	$\{2\},\{1,3\},\{1,5\},\{2,3\},\{4,5\}$
2	$\{1,2\},\{5,6\},\{1,3,5\}$
3	$\{2,5\},\{1,2,3\}$

Table 3.1: The values of α assigned to each simplex from the simplicial complex in Figure 2.4 for the purpose of a filtration.

value, α, associated with a simplex, σ, must be greater than the value associated with all of the faces of σ. That is a simplex cannot appear in the filtration until all of its faces have appeared in the filtration. We construct a filtration based on the simplicial complex in Figure 2.4 by applying the alpha values specified in Table 3.1.

We can see how the simplicial complex looks at each step of the filtration in Figure 3.1. The Betti numbers change throughout the filtration. For example $\beta_0 = 2$ at $\alpha = 0$ but $\beta_0 = 1$ where $\alpha \geq 1$ while $\beta_1 = 0$ at $\alpha = 0$, $\beta_1 = 2$ at $\alpha = 1$ and $\beta_1 = 3$ at $\alpha \geq 2$. Note that even though β_1 does not change between $\alpha = 2$ and $\alpha = 3$ the holes which contribute to this Betti number are not the same at those levels of the filtration. We can say that the hole represented by the 1-cycle $[1,2]-[1,3]+[2,3]$ is born at $\alpha = 2$ and dies at $\alpha = 3$. The difference between the filtration level of the death of a hole and the filtration level of the birth of a hole is its persistence, thus the hole represented by the 1-cycle $[1,2]-[1,3]+[2,3]$ has a persistence of 1.

Petri et al. used the concept of persistence to define a centrality measure on the edges of a simplicial complex [120]. The total persistence in the persistence homological scaffold of an edge is the amount of time it spends on the boundary of the various holes which appear and disappear throughout the filtration. The edge $\{1,5\}$ in the filtration in Table 3.1 has a total persistence of 4 because it is on the boundary of $[1,3]-[1,5]+[3,5]$ from $\alpha = 1$ to $\alpha = 2$ and on the boundary of $[1,5]-[1,6]-[4,5]+[4,6]$ which becomes $[1,5]-[1,6]+[5,6]$ from $\alpha = 1$ to the end of the filtration. The fact that it is necessary to pick a representative for each hole [120] means that this measure is problematic for the use of total persistence as a centrality measure and the necessity of having a filtration to work with also implies that it is not possible to apply it in many cases.



(a) (b) (c) (d)

Figure 3.1: Evolution of simplicial complex under filtration described in Table 3.1 for $\alpha = 0$ (a), $\alpha = 1$ (b), $\alpha = 2$ (c) and $\alpha = 3$ (d).

Through the use of simplicial complexes it is possible to extend the network-theoretic centralities to find the most important pair, triangle or tetrahedron, provided that the right centrality measure for the phenomenon and question under study are chosen. For instance Jiang and Omer [74] propose concepts of degree centrality, closeness centrality and betweenness centrality on simplicial complexes. They leverage Atkin's q–adjacency [9] which is that two simplices are q–adjacent if they share a common q–simplex. This definition means that for the simplicial complex in Figure 2.4 then $\{1,4,6\}$ is adjacent to $\{1,2\}$ because they share the common node $\{1\}$. Jiang and Omer defined a q–degree centrality of a simplex in a simplicial complex to be the number of other simplexes which share a common q–simplex with the original simplex. For the simplicial complex in Figure 2.4 we have that the 0–degree of $\{1,4,6\}$ would be 13 because it shares a common 0-simplex with $\{1\}$, $\{4\}$, $\{6\}$, $\{1,2\}$, $\{1,3\}$, $\{1,4\}$, $\{1,5\}$, $\{1,6\}$, $\{4,5\}$, $\{4,6\}$, $\{5,6\}$, $\{1,2,3\}$, and $\{1,3,5\}$. They similarly define a q–closeness centrality based on the reciprocal of the sum of the shortest paths based on the q-adjacency for each simplex to every other simplex in the simplicial complex and a q–betweenness centrality based on the proportion of these shortest paths which a simplex is a part of. Jiang and Omer applied their centralities to the use of parks to enable interconnectedness between neighbourhoods in Tel Aviv. Each neighbourhood was represented by a node and each park by a simplex which contained a node if the border of the neighbourhood represented by that node was less than 500m from the border of the park. This representation meant that some parks were represented

by simplices of dimension 6, which is slightly unnatural given that these parks and neighbourhoods are embedded in a two-dimensional space. There were two parks in particular, one of which was represented by a 6-simplex and another by a 5-simplex, which shared multiple neighbourhoods in common but were isolated from the rest of the parks at the level of the 2-simplices and higher. These two parks had a very high degree because of the size of the simplices despite not being located in an area where many neighbourhoods were connected by multiple parks.

The degree centralities introduced by Moore et al. [104] are preferable in this respect in that they root the degree firmly in the interactions that are happening at the same dimension. Moore et al. defined the total degree of a k-simplex to be the number of other k-simplices to which a k-simplex is lower adjacent added to the number of other k-simplices to which it is upper adjacent. The use of the upper adjacency as part of this definition still poses problems in terms of the interpretation of this measure though. We can consider co-authorship networks which Moore et al. [104] studied. Consider three authors a, b, c who have exactly one publication together which was as part of a collaboration of a total of 50 authors. Let another author from that collaboration be denoted by d, we have that $\{a, b, c\}$ is upper adjacent to $\{a, b, d\}$, $\{a, c, d\}$ and $\{b, c, d\}$. By Theorem 2.30 we have that $\{a, b, c\}$ is also lower adjacent to $\{a, b, d\}$, $\{a, c, d\}$ and $\{b, c, d\}$ which means that the degree of $\{a, b, c\}$ by this definition is 6 multiplied by the 47 other co-authors on that publication for a total of 282. Now consider the situation that the same three co-authors have published together but have also published in the following collaborations $\{a, b, d\}, \{a, b, e\}, \{a, b, f\}, \{a, c, g\}, \{a, c, h\}, \{b, c, i\}$. The degree of $\{a, b, c\}$ in this situation would be 6 despite those three authors being involved in more collaborations than in the first situation which is another example of why we have to be careful to ensure that the centrality measures we are using are relevant to the simplicial complex and to the question we wish to understand.

Bianconi and Rahmede proposed a different a generalised degree centrality for simplicial complexes in 2015 [17]. The generalised degree was defined on a special kind of simplicial complex called a Complex Quantum Network Manifold (CQNM) which, for a given level d, is constructed by gluing d-simplices together along the $(d-1)$ faces.

The generalised degree of a k-simplex, where $k < d$, is the number of d-simplices that the k-simplex is a face of. It is easy to see how this definition could be used for simplicial complexes which are not CQNMs by picking a level of the simplicial complex and defining the generalised degrees in relation to that level [16]. Raj and Bhattacharya recently built on the generalised degree by defining an adjacency on the d-simplices of a simplicial complex for each $k < d$ such that two d-simplices are adjacent if they share a common k-simplex [124] which is essentially a special case of the p-lower adjacency of [143] which we describe below. Raj and Bhattacharya used this adjacency to define shortest paths through the simplices which allowed them to create the generalised closeness and generalised betweenness [124].

Another network centrality measure which has been generalised to simplicial complexes is the PageRank by Schaub et al. [139] using the Hodge Laplacians that we discussed in Section 2.6. This centrality is the probability that a random walker on the k-simplices of a simplicial complex is at a given simplex at any given time.

Due to the existence of other notions of adjacency there are also other notions of walks on graphs which do not line up with the one we use here. For example, Mukherjee and Steenbergen's random walks [110] line up with the concept of lower adjacency that was discussed in Section 2.5.2.

Since the publication of [50] further work has been done in the realm of centralities on simplicial complexes. In 2020 Hernández-Serrano, Hernández-Serrano and Sánchez Gómez introduced a new simplicial degree [143] which was extended to new definitions of simplicial eigenvector centrality, simplicial betweenness, simplicial closeness and simplicial clustering coefficient [142]. They defined that a q-simplex, σ, and a r-simplex, τ, to be p-lower adjacent if they share a common p-face and to be p-upper adjacent if they are both faces of the same p-simplex. In the simplicial complex depicted in Figure 2.4 we have that $\{1, 2, 3\}$ would be 0-lower adjacent to $\{3, 5\}$ through the common node 3 while $\{1\}$ and $\{3, 5\}$ would be 2-upper adjacent because these simplices are both faces of $\{1, 3, 5\}$. These three authors then used these notions of adjacency to create several different types of degree centrality for simplices including the maximal simplicial degree which counts the number of largest possible simplices that a simplex is a face of and

adds the number of simplices that it is lower adjacent to provided that these simplices are not upper adjacent to the original simplex and also that a simplex and its faces are counted together. Returning to the simplicial complex from Figure 2.4 we have that $\{1,3\}$ has a maximal simplicial degree of 3 because it is a face of $\{1,2,3\}$ and $\{1,3,5\}$ and is 1-lower adjacent to $\{1,4,6\}$. Note that the 1-lower adjacency of $\{1,3\}$ to $\{1,4\}$ does not count towards the maximal adjacency because it is included as part of $\{1,4,6\}$.

Santos et al. extended the simplicial eigenvector centrality which was introduced by Hernández-Serrano and Sánchez Gómez [142] to the case of hypergraphs and applied it to functional MRI networks [138]. This use of hypergraphs points us to alternative models which could be used to establish centralities on interactions of more than two nodes. A hypergraph, like a network or simplicial complex, consists of a set, V, of nodes and a set, E, of subsets of the nodes [22, Section 1.1]. In a hypergraph the members of E, known as hyperedges, can be any size [22]. A hypergraph differs from a network in that hyperedges can contain more than two nodes and differs from a simplicial complex because the existence of a hyperedge does not imply that all subsets of it are also hyperedges.

We return to the Friends example from Figure 2.1 and Figure 2.4. In Friends while Chandler, Monica and Phoebe all lived together at one time at no point during the show do Chandler and Phoebe live together without Monica therefore the hyperedge $\{1,2,3\}$ is a member of E in a hypergraph representation of this system but $\{1,2\}$ is not. The set of hyperedges in this example is given by $E = \{\{1,3\},\{1,4\},\{1,6\},\{2,3\},\{2,5\},\{3,5\},\{4,5\},\{5,6\},\{1,2,3\},\{1,3,5\},\{1,4,6\}\}$. We can see that we have added $\{1,2,3\},\{1,3,5\}$ and $\{1,4,6\}$ and dropped $\{1,2\},\{1,5\}$ and $\{4,6\}$ compared to the networks case. In the simplicial complexes case we would be obliged to include $\{1,2\},\{1,5\}$ and $\{4,6\}$ because they are faces of larger simplices.

We have three different representations of the same system which gives rise to the question of which one is correct. Torres et al. [159] explored this question through publication co-authorship networks. A network is the most appropriate way to represent co-authorship if we only want to consider whether or not two authors have worked

together on the same paper. The network theoretic node degree tells us the number of different co-authors an author has. This number is not returned by either the simplicial complex or hypergraph node degrees applied by Torres et al. which count different communities of collaborators and different collaborations respectively [159]. A simplicial complex is the preferable representation if we want to ask whether or not a set of authors have worked together on a paper at some point. Under the maximal simplicial node degree [143] used by Torres et al. the number of maximal simplices a node is a member of is counted which means that if a node is involved in one big collaboration and then involved in several smaller collaborations which only contain members of the larger collaboration then the smaller collaborations will not contribute to increasing the node degree. We can infer that the node degree under the simplicial complexes case is counting the number of different collaboration communities that a node is a part of. A hypergraph is more accurate if we only want to consider whether a set of authors has produced a paper together with no other authors involved. Torres et al. constructed a dummy example where only the node degree on a hypergraph was able to capture that a node was involved in the most different collaborations because under the network and simplicial complex representations the other nodes had increased node degrees due to their involvement in couple of large collaborations [159]. Ultimately, there are many different factors which should be considered when deciding how to model a problem. These factors include but are not limited to the quality of the data available, what an interaction would mean in the model, how an interaction relates to potential subinteractions and the question being posed of the model [159]. These considerations mean that there may not be a correct answer to the question of which model should be used in every circumstance, however, the choice can affect the insights which can be drawn about the problem under consideration [159]. As a result in Chapter 4 and Chapter 5 we explore the potential of hypergraph representations to be used instead of simplicial complexes for the applications we consider.

There are many centrality measures for hypergraphs including versions of eigenvector centrality which are different from the one used by Santos et al. [8, 15, 138, 161], subhypergraph centrality [49] and betweenness centrality [158].

We propose a slightly different set of simplicial centralities which, while still derived from the Laplacian and concepts of upper and lower adjacency, seek to isolate and study the interactions at particular levels of dimensionality. We apply these centrality measures to protein–protein interaction networks in Chapter 4 and Random Geometric Graphs/Wireless Sensor Networks in Chapter 5.

3.2 Adjacency Matrices in Simplicial Complexes

The goal of this section is to combine Goldberg's concepts of upper and lower adjacency [66] that we discussed in Section 2.5.2 into a general adjacency matrix for a simplicial complex that allows us to define general centrality indices for these mathematical objects.

Definition 3.1. Let C be a simplicial complex and let $\sigma_i, \sigma_j \in C$ be two k-simplices. Then, for $k \geq 1$, σ_i and σ_j are considered **simplex adjacent** if they are lower adjacent and not upper adjacent. For $k = 0$ two simplices shall be simplex adjacent if they are upper adjacent. We use $\sigma_i \backsim \sigma_j$ to denote that two k-simplices, σ_i, σ_j are simplex adjacent.

The choice to define simplex adjacency as being lower adjacent and not upper adjacent avoids the issue of any centrality measures defined using this notion of adjacency being overpowered by the presence of large simplices in the way the degree centralities in [104] are. The simplex adjacency from Definition 3.1 does not completely ignore the effects of upper adjacency and uses it to isolate the interactions that can only happen between simplices of the same size. By using the upper adjacency in this way it avoids a large simplex boosting the centralities of its faces which would happen if only the lower adjacency were considered. Instead the interactions between these simplices are considered within the interactions of the larger simplex at the appropriate level. This notion of adjacency lines up nicely with the extensively studied higher order Laplacians of simplicial complexes [107] which we discussed in Section 2.6. When we compare Definition 3.1 and Definition 2.46 we can see that an off-diagonal entry of the higher order Laplacian matrix is non-zero if and only if the simplices represented by that row and

column are simplex adjacent.

However, an ideal definition of simplex adjacency would also be monotonic with respect to the addition of simplices. That is, we would expect that adding an additional simplex to a simplicial complex should increase the connectivity of the simplicial complex in that more simplices should become simplex adjacent to each other. We are aware that Definition 3.1 is non-monotonic because the addition of any k-simplex results in all of its faces becoming upper adjacent to each other and hence not simplex adjacent. The clearest example of this is the clique complex of any complete graph in which no two k-simplices for $k > 0$ are simplex adjacent to each other because every possible k-simplex is included in the simplicial complex and as a result they are all upper adjacent to each other. It is clearly not natural that the graph with the highest possible density of connections should give rise to the simplicial complex with the lowest possible density of simplex adjacent simplices.

We believe that the advantages of Definition 3.1, in that it prevents the presence of a couple of large simplices from drowning out the effects of the rest of the simplicial complex and it respects the Hodge Laplacian, are worth the sacrifice of its non-monotonicity. Additionally, we were not able to find a definition of simplex adjacency that had these advantages and was also monotonic.

Given the notion of simplex adjacency and its relation to the Hodge Laplacian it is natural to define a simplex adjacency matrix for each level of the simplicial complex.

Definition 3.2. Let C be a simplicial complex and let $\sigma_i, \sigma_j \in C$ be two k-simplices. Then, for $k \geq 1$ the **simplex adjacency matrix** A_k at the k-level in the simplicial complex has entries defined by

$$(A_k)_{ij} = \begin{cases} 1 & \text{if } \sigma_i \smile \sigma_j \text{ and } \sigma_i \nsucc \sigma_j; \\ 0 & \text{if } i = j \text{ or } \sigma_i \nsim \sigma_j \text{ or } \sigma_i \nsucc \sigma_j; \end{cases}$$

for $k = 0$ the simplex adjacency matrix is given by the network theoretic adjacency matrix.

For example A_1 and A_2 for the simplicial complex in Figure 2.4 are given by

$$A_1 = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{matrix} \{1,2\} \\ \{1,3\} \\ \{1,4\} \\ \{1,5\} \\ \{1,6\} \\ \{2,3\} \\ \{2,5\} \\ \{3,5\} \\ \{4,5\} \\ \{4,6\} \\ \{5,6\} \end{matrix} \tag{3.1}$$

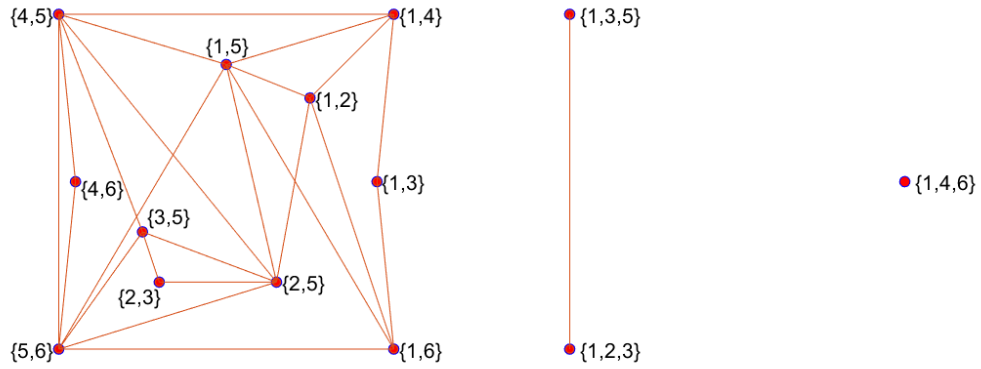
$$A_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{matrix} \{1,2,3\} \\ \{1,3,5\} \\ \{1,4,6\} \end{matrix} \tag{3.2}$$

Definition 3.3. The **underlying network of simplices** for the k-simplices of a simplicial complex is constructed by mapping every k-simplex in the simplicial complex to a node in the underlying network. Two nodes in the underlying network are adjacent to each other if and only if the two simplices they represent are simplex adjacent.

The underlying network of simplices for the 1-simplices and 2-simplices of the simplicial complex in Figure 2.4 are shown in Figure 3.2(a) and (b) respectively. The adjacency matrices of these underlying networks of simplices are identical to A_1 and A_2 in Equations (3.1) and (3.2) respectively.

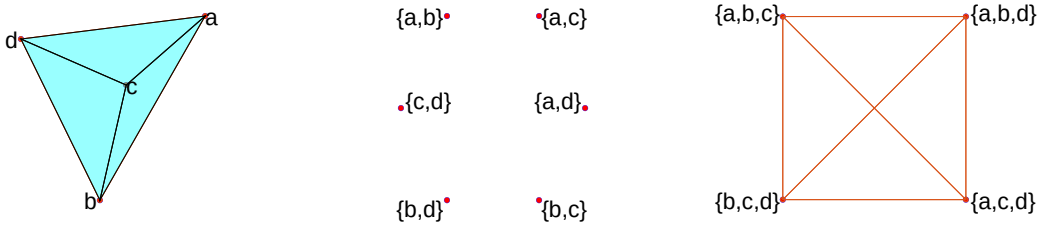
Note that this construction is similar to the concept of the conjugate complex introduced by Atkin [9] except that it is specific to a particular dimension level and a simplex can be included even if it is a face of a larger simplex.

We can use the underlying network of simplices to demonstrate the non-monotonicity of simplex adjacency by comparing Figure 3.3 and Figure 3.4 which has an additional simplex, $\{a, b, c, d\}$, and shows the case of the clique complex on a complete graph.



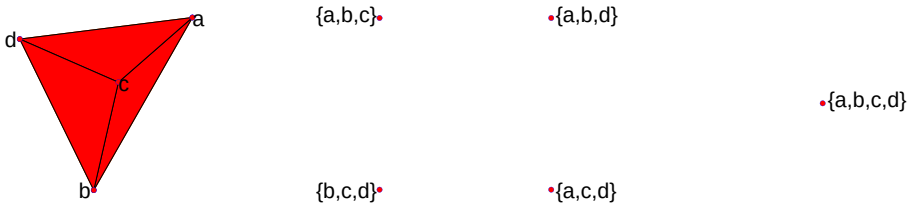
(a) (b)

Figure 3.2: The Underlying Network of Simplices for the Friends simplicial complex for the 1-simplices (a) and the 2-simplices (b).



(a) (b) (c)

Figure 3.3: A simplicial complex consisting of four 2-simplices, $\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\}$ and all of their faces (a), its underlying network of 1-simplices (b) and underlying network of 1-simplices (c).



(a) (b) (c)

Figure 3.4: A simplicial complex consisting of a 3-simplex, $\{a, b, c, d\}$ and all of its faces (a), its underlying network of 2-simplices (b) and underlying network of 3-simplices (c).

3.3 Simplicial Shortest Path Distance

In this section we extend the concept of shortest path distance to the different levels of a simplicial complex. We start by extending the concept of walks to simplicial complexes.

Definition 3.4. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $k \geq 1$, $\sigma_i \in R_k$ and $\zeta_i \in R_{k-1}$. Then, a **k-walk** is a sequence of alternating k-simplices and $(k-1)$-simplices $\sigma_1, \zeta_1, \sigma_2, \zeta_2, \ldots, \zeta_{r-1}, \sigma_r$ such that for each $i \in \{1, \ldots, r-1\}$ σ_i and σ_{i+1} are simplex adjacent and ζ_i is a face of both σ_i and σ_{i+1}. For $k = 0$ a walk on the 0-simplices is just a walk in the normal graph-theoretic sense.

On the simplicial complex from Figure 2.5, we have that $\{1,3,4\}, \{3,4\}, \{3,4,5\}$, $\{4,5\}, \{4,5,6\}, \{4,5\}, \{3,4,5\}, \{3,4\}, \{2,3,4\}$ is a 2-walk. Meanwhile, $\{6,9\}, \{6\}, \{6,7\}, \{6\}$, $\{5,6\}, \{5\}, \{3,5\}, \{3\}, \{2,3\}$ is a 1-walk.

Definition 3.5. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma_i \in R_k$ and $\zeta_i \in R_{k-1}$. A **k-shortest path** between two k-simplices $\sigma_a, \sigma_b \in R_k$ is a k-walk, $\sigma_a, \zeta_1, \sigma_2, \zeta_2, \ldots, \sigma_n, \zeta_n, \sigma_b$, such that n is minimized. The value n is the **k-shortest path length** between the two k-simplices σ_a, σ_b. We denote this value $d_k(\sigma_a, \sigma_b) = n$. If there is no path between σ_a and σ_b then $d_k(\sigma_a, \sigma_b) = \infty$.

Definition 3.6. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R. A simplicial complex is **k-connected** if and only if there does not exist a pair of k-simplices $\sigma_a, \sigma_b \in R_k$ such that $d_k(\sigma_a, \sigma_b) = \infty$, i.e for every pair of k-simplices $d_k(\sigma_a, \sigma_b)$ is finite.

Note that a simplicial complex being k-connected does not mean that it is $(k+1)$-connected or $(k-1)$-connected. The simplicial complex in Figure 2.5 is 0-connected but not 1-connected because $\{1,2\}$ and $\{7,8\}$ are not simplex adjacent to any of the other 1-simplices. Many of the real world networks we shall introduce in later sections are 1-connected but not 2-connected. In addition, a simplicial complex from the family S_l^k (See Figure 2.6), is k-connected but it is not $(k-1)$-connected. The central $(k-1)$-simplex is upper adjacent to every other $(k-1)$-simplex and hence is not simplex adjacent to any of them.

Definition 3.7. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R. A set of k-simplices $S_k \subseteq R_k$ is a **k-connected component** if it is the case that for any two k-simplices $\sigma_a, \sigma_b \in S_k$ we have $d_k(\sigma_a, \sigma_b) < \infty$ and for any $\sigma_c \in S_k$ and $\sigma_d \in R_k \setminus S_k$ we have that $d_k(\sigma_c, \sigma_d) = \infty$.

Theorem 3.8. *Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma_i \in R_k$ and $\zeta_i \in R_{k-1}$. Let R be k-connected then the simplicial shortest path length between two k-simplices is an extended metric.*

Proof. By definition $d_k(\sigma_a, \sigma_b) \geq 0$ for all $\sigma_a, \sigma_b \in R_k$. Clearly $d_k(\sigma_a, \sigma_b) = 0 \iff \sigma_a = \sigma_b$.

To prove $d_k(\sigma_a, \sigma_b) = d_k(\sigma_b, \sigma_a)$ assume $d_k(\sigma_a, \sigma_b) = n$. Then the k-shortest path from σ_a to σ_b is of the form $\sigma_a, \zeta_1, \sigma_2, \zeta_2, \ldots, \sigma_{n-1}, \zeta_{n-1}, \sigma_n, \zeta_n, \sigma_b$ which means that there is a k-walk from σ_b to σ_a of the form $\sigma_b, \zeta_n, \sigma_n, \zeta_{n-1}, \sigma_{n-1}, \ldots, \zeta_2, \sigma_2, \zeta_1, \sigma_a$. We can then relabel $\zeta_1 \to \zeta_n, \sigma_2 \to \sigma_n, \zeta_2 \to \zeta_{n-1}, \sigma_3 \to \sigma_{n-1}, \ldots, \zeta_n \to \zeta_1$ and so on to give a k-walk from σ_b to σ_a of the form $\sigma_b, \zeta_1, \sigma_2, \zeta_2, \ldots, \sigma_{n-1}, \zeta_{n-1}, \sigma_n, \zeta_n, \sigma_a$ thus $d_k(\sigma_b, \sigma_a) \leq n$. Let there be a k-walk shorter than length n from σ_b to σ_a then we could relabel it to create a k-walk shorter than length n from σ_a to σ_b which would contradict $d_k(\sigma_a, \sigma_b) = n$. Thus $d_k(\sigma_b, \sigma_a) = n$ and $d_k(\sigma_a, \sigma_b) = d_k(\sigma_b, \sigma_a)$.

To prove $d_k(\sigma_a, \sigma_c) \leq d_k(\sigma_a, \sigma_b) + d_k(\sigma_b, \sigma_c)$ let $d_k(\sigma_a, \sigma_b) = n$ and $d_k(\sigma_b, \sigma_c) = m$ then there is a k-walk from σ_a to σ_b of the form $\sigma_a, \zeta_1, \sigma_2, \zeta_2, \ldots, \sigma_{n-1}, \zeta_{n-1}, \sigma_n, \zeta_n, \sigma_b$ and k-walk from σ_b to σ_c of the form $\sigma_b, \zeta_1, \sigma_2, \zeta_2, \ldots, \sigma_{m-1}, \zeta_{m-1}, \sigma_m, \zeta_m, \sigma_c$ we can combine these k-walks and relabel the simplices in the second walk by the rules $\sigma_b \to \sigma_{n+1}, \zeta_i \to \zeta_{n+i}, \sigma_i \to \sigma_{n+i}$ to form a k-walk from σ_a to σ_c of the form $\sigma_a, \zeta_1, \sigma_2, \zeta_2, \ldots, \sigma_{n-1}, \zeta_{n-1}, \sigma_n, \zeta_n, \sigma_{n+1}, \zeta_{n+1}, \sigma_{n+2}, \zeta_{n+2}, \ldots, \sigma_{n+m-1}, \zeta_{n+m-1}, \sigma_{n+m}, \zeta_{n+m}, \sigma_c$ which implies that $d_k(\sigma_a, \sigma_c) \leq n + m = d_k(\sigma_a, \sigma_b) + d_k(\sigma_b, \sigma_c)$. □

For instance, on the simplicial complex from Figure 2.5, we have that $\{1,3,4\}, \{3,4\}, \{3,4,5\}, \{3,4\}, \{2,3,4\}$ is a 2-shortest path from $\{1,3,4\}$ to $\{2,3,4\}$ and we have $d_2(\{1,3,4\}, \{2,3,4\}) = 2$. Meanwhile, $\{2,4\}, \{4\}, \{4,6\}, \{6\}, \{6,7\}$ is a 1-shortest path between $\{2,4\}$ and $\{6,7\}$ and we have $d_1(\{2,4\}, \{6,7\}) = 2$.

Definition 3.9. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma \in R_k$. The **k-eccentricity** $\epsilon_k(\sigma)$ is the largest k-shortest path

distance between σ and any other k-simplex.

Definition 3.10. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma \in R_k$. The **k-diameter**, D_k of a simplicial complex is the maximum k-eccentricity of any $\sigma \in R_k$. That is $D_k = \max_{\sigma \in R_k} \epsilon_k(\sigma)$.

As an example, in the simplicial complex $t^2(1,2,4)$, depicted in Figure 2.6, the central 2-simplex has 2-eccentricity 1 because it is simplex adjacent to all the other 2-simplices in the complex. However, all the peripheral 2-simplices have a 2-eccentricity of 2 because the shortest path from a peripheral 2-simplex on one arm to a peripheral 2-simplex on another is through the central 2-simplex for a shortest path of length 2 which means that $t^2(1,2,4)$ has 2-diameter 2.

Given a notion of shortest path distance we are now equipped to define the average simplicial shortest path distance.

Definition 3.11. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma_i, \sigma_j \in R_k$. Let R be k-connected then the **k-average simplicial shortest path length**, l_k, is the average k-shortest path distance for all possible pairs of k-simplices in the network. It is given by

$$l_k = \frac{2\sum_{i<j} d_k(\sigma_i, \sigma_j)}{\|R_k\|(\|R_k\| - 1)}. \tag{3.3}$$

Note that if the simplicial complex is not k-connected, we can analyse each k-connected component separately.

We now prove bounds on the k-average path length.

Theorem 3.12. *A sharp lower bound on l_k is* 1.

Proof. For l_k to be less than 1 there would need to be two distinct k-simplices, σ_i, σ_j such that $d_k(\sigma_i, \sigma_j) < 1$ which would imply $d_k(\sigma_i, \sigma_j) = 0$ and hence $\sigma_i = \sigma_j$ by the properties of an extended metric.

The lower bound $l_k = 1$ is achieved by a simplicial complex of the form S_r^k which is easy to check. A simplicial complex of the form S_r^k consists of a $(k-1)$-simplex $\{1, 2, \ldots, k\}$ and some k-simplices of the form $\{1, 2, \ldots, k, i\}$, where $i > k$, in addition

to all subsimplices necessary by the closure axiom. Hence, all k-simplices are lower adjacent to each other by the $(k-1)$-simplex $\{1, 2, \ldots, k\}$ and they are not upper adjacent to each other because there are no $(k+1)$-simplices. Thus, every k-simplex is simplex adjacent to every other k-simplex and the k-shortest path distance between any two k-simplices is 1. Hence, the k-average path length is 1 and the lower bound of l_k is achieved. □

A general upper bound of l_k is impossible to establish due to the dependence on the number of simplices, $\|R_k\|$. However, if we fix both k and $\|R_k\|$ then we can prove the following result.

Theorem 3.13. *Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma_i, \sigma_j \in R_k$. Let R be k-connected and let $\|R_k\| \geq 2$ be the fixed number of k-simplices in R. Then, an upper bound of l_k is*

$$\frac{\|R_k\| + 1}{3}. \tag{3.4}$$

Proof. If $\|R_k\| = 2$ then $\sum_{i<j} d_k(\sigma_i, \sigma_j) = 1$, the simplicial complex is k-connected and there are only 2 k-simplices hence they must be simplex adjacent. Thus

$$l_k = \frac{2\sum_{i<j} d_k(\sigma_i, \sigma_j)}{\|R_k\|(\|R_k\| - 1)} = 1. \tag{3.5}$$

In addition $\frac{\|R_k\|+1}{3} = 1$. Hence the theorem holds for $\|R_k\| = 2$.

Assume that the theorem holds for $\|R_k\| \leq n$. Let $\|R_k\| = n + 1$. Then to maximize l_k we need to maximize $\sum_{i<j} d_k(\sigma_i, \sigma_j)$. Pick a k-simplex σ_1. First, we maximize $\sum_j d_k(\sigma_1, \sigma_j)$. For $d_k(\sigma_1, \sigma_j) = y$ for some $\sigma_j \in R_k$, it must be the case that $d_k(\sigma_1, \sigma_m) = y - 1$ for some $\sigma_m \in R_k$ such that $\sigma_m \backsim \sigma_j$ which means that the largest possible value of $d_k(\sigma_1, \sigma_j)$ for some $\sigma_j \in R_k$ is $\|R_k\| - 1 = n$. We can evaluate $\max \sum_j d_k(\sigma_1, \sigma_j) = (\|R_k\|-1)+(\|R_k\|-2)+\cdots+1 = T_{\|R_k\|-1} = T_n$ where T_z represents the z^{th} triangle number which implies that $d_k(\sigma_a, \sigma_1) = 1$ for precisely one k-simplex, $\sigma_a \in R_k$. If $d_k(\sigma_a, \sigma_1) = 1$ for precisely one k-simplex, $\sigma_a \in R_k$ then σ_1 is simplex adjacent to precisely one other k-simplex, namely σ_a. Because σ_1 is simplex adjacent to only one other simplex, σ_1 can be

removed without affecting the k-shortest path distances between any other k-simplices. We now have a simplicial complex such that $\|R_k\| = n$. We know that the upper bound of the k-average path distance for this smaller simplicial complex is $\frac{\|R_k\|+1}{3} = \frac{n+1}{3}$ by assumption where $\frac{(n-1)n(n+1)}{6}$ is the contribution given by $\sum_{i<j} d_k(\sigma_i, \sigma_j)$. We also know that the largest number we can add to the sum $\sum_{i<j} d_k(\sigma_i, \sigma_j)$ by the addition of a k-simplex is given by $T_n = \frac{n(n+1)}{2}$. Thus when $\|R_k\| = n+1$ we have

$$\max\left(\sum_{i<j} d_k(\sigma_i, \sigma_j)\right) = \frac{(n-1)n(n+1)}{6} + \frac{n(n+1)}{2} = \frac{n(n+1)(n+2)}{6}. \tag{3.6}$$

Equation (3.6) allows us to evaluate the upper bound of l_k as

$$\frac{\frac{(\|R_k\|-1)\|R_k\|(\|R_k\|+1)}{3}}{(\|R_k\|-1)\|R_k\|} = \frac{\|R_k\|+1}{3}. \tag{3.7}$$

A simplicial complex of the form P_r^k achieves this bound. □

3.4 Centralities Based on Simplicial Shortest-Path

We are now in a position to generalize some centrality notions for simplices which are based on the simplicial shortest path distance. The simplest of all centrality measures is the degree.

Definition 3.14. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma \in R_k$. The **k-degree**, $\delta_k(\sigma)$ is the number of other k-simplices to which σ is simplex adjacent.

Note that we drop k from the degree where it is clear at which level of the simplicial complex we are working as we do for the other simplicial centrality measures that are discussed in this thesis. We will now relate the degrees of the 1- simplices to the degrees of the 0-simplices.

Theorem 3.15. *Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma = \{i, j\} \in R_1$. The degree of σ can be calculated by the formula*

$$\delta_1(\sigma) = \delta_0(i) + \delta_0(j) - (2 + 2T) \tag{3.8}$$

where T is the number of 2-simplices of which σ is a face.

Proof. Recall that a pair of 1-simplices are said to be simplex adjacent if they are lower adjacent and not upper adjacent. That is they share a common 0-simplex and are not faces of the same 2-simplex. Let's assume, for now, that there are no 2-simplices in the simplicial complex so that we can focus on the contribution from the lower adjacencies. There are two ways in which σ can be lower adjacent to other 1-simplices, either through i being a shared common 0-simplex or through j. The number of 1-simplices which are lower adjacent to σ through i is given by $(\delta_0(i) - 1)$ because i is also adjacent to j at the level of the 0-simplices. Similarly, the number of 1-simplices which are lower adjacent to σ through j is given by $(\delta_0(j) - 1)$ thus in the event that there are no 2 simplices then $\delta_1(\sigma) = \delta_0(i) + \delta_0(j) - 2$.

Now let $\zeta = \{i, j, m\}$ be a 2-simplex of which σ is a face. The other two faces of ζ, which are $\{i, m\}$ and $\{j, m\}$, are lower adjacent to σ but they are now both upper adjacent to σ and therefore not simplex adjacent to it. We have to subtract 2 from the value of $\delta_1(\sigma)$ to account for the effect of this upper adjacency. The 2-simplex, ζ, was chosen arbitrarily and so for each 2-simplex of which σ is a face there are two 1-simplices which are lower adjacent to σ are also be upper adjacent to it and therefore are not simplex adjacent to σ which leads us to the stated equation. □

Definition 3.16. Let a simplicial complex, R, be given. If $p(\delta_k)$ is the probability of finding a k-simplex of degree δ_k in R, then the **k-degree distribution** of the k-simplices is the probability distribution of the degrees of the k-simplices across the whole of the simplicial complex.

The degrees of the 1-simplices of the Friends simplicial complex depicted in Figure 2.4 are displayed in Tables 3.2 and 3.3 along with the simplicial closeness and simplicial subgraph centralities at this level which shall be introduced in the rest of this chapter. We can see that there are 4 1-simplices which have the highest degree which are $\{1, 5\}, \{2, 5\}, \{4, 5\}, \{5, 6\}$. Three of these four 1-simplices are not faces of any 2-simplices which suggests that 1-simplices with high degree are likely to not be faces of many 2-simplices. In Chapter 5 we will look at wireless sensor networks where this

	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{1,5\}$	$\{1,6\}$	$\{2,3\}$
δ_1	4	2	4	6	4	2
Simplicial Closeness	0.59	0.43	0.59	0.71	0.59	0.43
Simplicial Subgraph Centrality	10.80	3.24	9.39	21.35	9.39	4.28

Table 3.2: The results of applying various simplicial centrality measures to the 1-simplices of the Friends simplicial complex depicted in Figure 2.4. This table is continued in Table 3.3.

	$\{2,5\}$	$\{3,5\}$	$\{4,5\}$	$\{4,6\}$	$\{5,6\}$
δ_1	6	4	6	2	6
Simplicial Closeness	0.67	0.59	0.71	0.48	0.71
Simplicial Subgraph Centrality	20.20	11.38	20.14	4.88	20.14

Table 3.3: Continued from Table 3.2: The results of applying various simplicial centrality measures to the 1-simplices of the Friends simplicial complex depicted in Figure 2.4.

property will be useful to detect regions of an area being monitored where the loss of a sensor will lead to a gap in the coverage.

All four of the 1-simplices with the highest simplicial degree contain node 5 which demonstrates that 1-simplices with high degree are likely to be found around nodes with high degree which are not members of many 2-simplices. We could therefore aggregate the degrees of the 1-simplices to find nodes which are faces of many 1-simplices but few 2-simplices to find the nodes which have such structures. Table 3.4 displays the mean degree of the 1-simplices for each node in the Friends simplicial complex. This table shows that nodes 1 and 5 which had identical results for all of the network based centrality measures in Section 2.2 are separated by taking the mean of the results for δ_1. We aggregate the centralities of the 2-simplices in order to find essential proteins in Chapter 4.

There are other methods to aggregate the centrality measures for the higher order simplices down to node level metrics such as counting the number of occasions where a node appears in the top x simplices or the top $x\%$ of simplices in a ranking. This method is the one we will use to reduce the centrality measures of the 1-simplices to figure out which nodes in a wireless sensor network are most likely to induce gaps in the coverage if they are removed in Chapter 5.

	1	2	3	4	5	6
Average δ_1	4.00	4.00	2.67	4.00	5.60	4.00
Average Simplicial Closeness	0.58	0.56	0.49	0.59	0.68	0.59
Average Simplicial Subgraph Centrality	10.84	11.76	6.30	11.47	18.64	11.47

Table 3.4: The average of the simplicial centralities from Table 3.2 and 3.3 of the 1-simplices of which each node is a member.

3.4.1 Simplicial Closeness

Here we generalize the concept of closeness centrality from Definition 2.16 to simplicial complexes.

Definition 3.17. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma_i, \sigma_j \in R_k$. Let R be k-connected then the **simplicial farness** of σ_i is the sum of its k-shortest path distances to all other k-simplices, $\sum_{i \neq j} d_k(\sigma_i, \sigma_j)$.

The **simplicial closeness** is the reciprocal of simplicial farness. That is

$$CC_k(\sigma_i) = \frac{\|R_k\| - 1}{\sum_{i \neq j} d_k(\sigma_i, \sigma_j)} \tag{3.9}$$

where $(\|R_k\| - 1)$ is a normalization factor.

If the simplicial complex is not k-connected then $\sum_{i \neq j} d_k(\sigma_i, \sigma_j)$ would be considered undefined or ∞ for all k-simplices in the simplicial complex. In this case we can calculate simplicial harmonic closeness instead which is a generalization of a network-theoretic definition that can be found in [130].

Definition 3.18. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma_i, \sigma_j \in R_k$ then the simplicial harmonic closeness of a k-simplex σ_i is defined as follows

$$HC_k(\sigma_i) = \sum_{i \neq j} \frac{1}{d_k(\sigma_i, \sigma_j)}, \tag{3.10}$$

where we treat $\frac{1}{\infty} = 0$.

Lemma 3.19. *The upper bound of the normalized simplicial closeness centrality is* 1 *which can be attained by all simplices in a simplicial complex of the form* S^k_l.

Proof. In a simplicial complex of the form S_l^k we have $l = \|R_k\|$ k-simplices which are all adjacent to each other. Thus if we select a particular k-simplex $\sigma_i \in R_k$ we have that $d_k(\sigma_i, \sigma_j) = 1$ for all $\sigma_j \in R_k$ such that $\sigma_i \neq \sigma_j$ which gives $\sum_{\sigma_i \neq \sigma_j} d_k(\sigma_i, \sigma_j) = \|R_k\| - 1$. Hence $CC_k(\sigma_i) = \frac{\|R_k\|-1}{\|R_k\|-1} = 1$. □

The peripheral 2-simplex which is only simplex adjacent to the central 2-simplex in the complex $t^2(1,2,4)$ has simplicial closeness given by $\frac{7}{13}$. We have $\sum_{\sigma \neq \sigma_j} d_k(\sigma, \sigma_j) = 1+2+2+2+2+2+2 = 13$ where σ is the given simplex and σ_j is a run through of the other simplices. The 1 is contributed by the shortest path from σ to the central simplex while the 2s are given by the shortest path distances from σ to the other peripheral simplices on the other branches. We also have that $\|R\| - 1 = 7$ for the normalization.

To give an example from the simplicial complex in Figure 2.5 we need to use the definition of harmonic simplicial closeness. So to calculate the simplicial closeness of $\{2,3,4\}$ we have $HC_k(\{2,3,4\}) = \frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{\infty} + \frac{1}{\infty} + \frac{1}{\infty} = 2$ because it is simplex adjacent to $\{3,4,5\}$ and has shortest path distance 2 to both $\{1,3,4\}$ and $\{4,5,6\}$. There is no k-path from $\{2,3,4\}$ to any of the other simplices.

Furthermore, the closeness centralities of the 1-simplices from the Friends simplicial complex from Figure 2.4 are displayed in Tables 3.2 and 3.3 and they are reduced to the average results for the nodes in Table 3.4. We can see that the edges which contain node 5 have higher closeness than their counterparts which contain node 1. The reduced number of 2-simplices which node 5 is a part of compared to node 1 means that node 5's edges have more direct access to each other which reduces the length of the 2-shortest paths between them. This difference is also present when the centralities are aggregated to the node level. In both cases the simplicial closeness offers greater granularity than the simplicial degree.

3.4.2 Simplicial Subgraph Centrality

We now move to the concepts of centrality based on taking powers of the adjacency matrices of simplicial complexes.

Theorem 3.20. *Let A_k be the simplex adjacency matrix between k-simplices in a sim-*

plicial complex. Then, $(A_k)_{ij}^m$ *gives the number of* k*-walks of length* m *between* k*-simplex,* σ_i *and* k*-simplex,* σ_j*.*

Proof. Every walk on the underlying network of simplices for a given simplex size k, has a corresponding k-walk over the k-simplices. We have that A_k is also the adjacency matrix for the nodes in the underlying network of simplices. Thus, powers of the simplex adjacency matrix can be used to give the numbers of walks of a given length on the underlying network of simplices. In particular, $(A_k)_{ij}^m = b$ means that there are b walks of length m between node i and node j in the underlying network of simplices at the k-simplex level which precisely corresponds to the existence of b k-walks of length m between simplex σ_i and simplex σ_j. □

We make a generalization of the exponential of the simplex adjacency matrix of k-simplices which relies on results from [48]. The following power series of the simplex adjacency matrix of k-simplices A_k in a simplicial complex converges to the corresponding matrix exponential

$$\sum_{l=0}^{\infty} \frac{A_k^l}{l!} = \exp(A_k). \tag{3.11}$$

Equation (3.11) is a direct generalization made possible by the adjacency matrices at the different levels of the simplicial complex.

Definition 3.21. Let a simplicial complex, R, be given and let R_k be the set of k-simplices of R with $\sigma_i \in R_k$ the k-simplex which corresponds to the i^{th} row of A_k. The **simplicial subgraph centrality** of σ_i, is given by

$$SC_k = (\exp(A_k))_{ii}. \tag{3.12}$$

For the simplicial complex in Figure 2.5 we have that the simplicial subgraph centrality of the 1-simplex $\{1, 4\}$ is 2.714. Note that any bounds on subgraph centrality for networks still hold due to the underlying network of simplices.

If we consider Tables 3.2 and 3.3 we can see the simplicial subgraph centralities for the Friends simplicial complex from Figure 2.4. The node average of these centralities for each node is displayed in Table 3.4. We can see that the highest ranked 1-simplex is $\{1, 5\}$. This time both nodes which belong to this edge have high node centralities and the edge is only involved in one 2-simplex and so there are a lot of small cycles involving this edge which boosts its simplicial subgraph centrality.

We can compare all three centralities which have been aggregated to the node level to the node-based centralities. In the case of the node centralities each of them had a tie for the most central node between nodes 5 and node 1 while the other four nodes were also all tied. When we look at the case of the 1-simplices those ties are broken which demonstrates that examining these extra levels of connection can lead to greater understanding of a system. The next two chapters demonstrate this effect in real-world situations.

Chapter 4

Analysis of Protein–Protein Interaction Networks

This chapter has been published as part of [50] with the exception that the interpretation of the degree distributions has been updated since that work. The idea to apply simplicial centrality measures to protein–protein interaction networks, to study their simplicial degree distributions, compare the simplicial centralities at different levels and search for essential proteins were contributions of Estrada to [50]. All calculations in this chapter were contributions of the author to [50]. The interpretation of the difference between the correlations of the centralities at the same level compared to centralities at different levels was a contribution of the author while Estrada identified the necessity of using the simplicial centralities to create a node based ranking for the purposes of identifying essential proteins and the author suggested the method of taking the mean of the simplices that each node was involved in. The author contributed the explanation of why an edge between two high degree nodes would be unlikely to have a high edge degree while Estrada contributed the comparison between an edge with two medium degree nodes and an edge between a low degree node and a high degree node. The interpretation of why the triangle centrality measures were good at picking out essential proteins was a contribution by Estrada. Both Estrada and the author reviewed all aspects of [50].

4.1 Literature Review

Proteins are incredibly important because of the number of roles they play in the functioning of cells. Among these roles, proteins can be used as catalysts in biological systems (i.e. enzymes). Enzymes have a variety of industrial applications including stain removal in detergents, aiding the digestibility of animal feed and maturing alcoholic beverages [86]. Another important biological role of proteins is their use by muscles where interactions between proteins result in a muscle contraction [151]. These roles are rarely performed by individual proteins instead they are caused by groups of proteins acting collaboratively [168]. When groups of proteins operate in this way it is known as a protein–protein interaction [168]. Within a cell a protein can be involved in several different interactions with many other proteins in combinations of large or small numbers [168].

In protein–protein interaction networks the function of the proteins within the cell define the structure of the network [125]. That is the proteins are the nodes and an edge exists between two nodes if their proteins take part in performing the same function for the cell. Three or more proteins can be involved in performing a function and so it is natural to consider extending the protein–protein interaction networks to protein–protein interaction simplicial complexes [98, 148] or protein–protein interaction hypergraphs [87, 92, 95]. Taking the simplicial complexes approach means that every simplex would signify that its members were part of a set of proteins which performed a function in the cell. This interpretation allows the possibility that 50 proteins perform a function together and four proteins which were part of that interaction but no other interactions would be considered to be a 3-simplex and have the same weight within the simplicial complex as 4 proteins which performed a function as part of a 4-way interaction. Therefore, a hypergraphs approach provides a more accurate model of the functional interactions between the proteins than the simplicial complexes one because every hyperedge represents the proteins that carried out a function.

However, the simplicial complexes framework gives us access to tools which are not available to hypergraphs. Song applied a persistent homology approach to protein–protein interactions on the Human PPI network and used a filtration based on the

p-value of each set of associated genomes to introduce the nodes of the PPI network and their interactions into the simplicial complex [148]. Song found 18 1-dimensional holes which persisted throughout the filtration and some of the proteins which were situated on the hole that persisted longest were associated with cancer and ageing which are known to be linked [170].

Consider the case that there are four proteins which are involved together in multiple interactions with no fifth protein also involved in all of the same interactions and that these four proteins do not perform a function together on their own. Then these four proteins would not form a hyperedge in the hypergraph representation and therefore would not appear to be as important to the functioning of the cell in this case. In the simplicial complexes case these four nodes would form a 3-simplex which would be at a cross-roads of many other simplices. Here we will consider protein–protein interactions from a simplicial complexes perspective and demonstrate that this perspective leads to insights about which proteins in a cell are essential.

Unfortunately, the methods by which we can detect which proteins are interacting with each other are fairly imprecise [168]. Yeast Two–Hybrid is an approach to identifying protein–protein interactions where one protein of interest is attached to a DNA binding domain and the other is attached to an activation domain. They are then brought closer together to see if they will interact [165]. This approach has a draw-back from our perspective which is that it can only assess one pair of proteins at a time; additionally not all proteins are appropriate for treatment in this way since the protein which is attached to the DNA binding domain must be capable of initiating transcription [165] and the fact that proteins are attached to a DNA binding domain or an activation domain in order to carry out this experiment may alter the results [165]. Von Mering et al. [168] found that when they filtered the data the Yeast Two–Hybrid method had an accuracy of 10%, which was third best of the six methods they studied, compared to a reference set of interactions in yeast but covered less than 1% of the possible interactions.

Another method of assessing the existence of interactions between proteins is Mass Spectroscopy of purified complexes of which there are two varieties which are Tandem

Affinity Purification (TAP) and High Throughput Mass-Spectrometric Protein Complex Identification (HMS-PCI) [168]. TAP works by selecting a protein of interest and attaching another molecule to it. This protein is then fished out of the cell along with the other proteins that it has interacted with. The nature of these proteins and their interactions is then determined by mass spectrometry [156]. According to Von Mering et al. [168] TAP has a similar accuracy to the Yeast Two–Hybrid method and covered more of the reference data set than the other six methods they considered. HMS-PCI follows more or less the same process but was reported to be less accurate and to cover fewer interactions than TAP [168].

Synthetic Lethality is a third technique for detecting protein–protein interactions. It describes a relationship between genes rather than proteins and any proteins which are encoded by those genes are said to interact if the two genes are synthetically lethal [114]. Two genes, a, b are synthetically lethal if there exists a non-lethal mutation in a and a non-lethal mutation in b such that the combination of these two mutations causes the inviability of the cell [114]. It has been suggested that where two genes are synthetically lethal the two sets of proteins they encode are essentially performing the same functions independently [114], meaning that there is inherently a redundancy written into the cell. Of all of the individual methods that Von Mering et al. [168] considered Synthetic Lethality had the highest accuracy of confirmed interactions versus their reference data set but had a lower coverage compared to all other methods except Yeast Two–Hybrid.

The penultimate method of protein interaction detection Von Mering et al. [168] considered are In-Silico detections which are essentially computer simulations to determine whether protein–protein interactions have happened or not. These detections involve examining the genome of an organism to search for indicators that the proteins interact [168]. There are three such indicators, whether or not the proteins are encoded by conserved operons, whether they are present or absent together in the fully sequenced genome and whether they have been found fused into one polypeptide chain [168]. According to Von Mering et al. [168], In-Silico predictions are only slightly less accurate than Synthetic lethality, TAP and Yeast Two–Hybrid, they cover around 10% of the interactions with the true value varying based on choice of parameters.

The final protein interaction detection method we look at here is Correlated mRNA expression. In this method a cell is exposed to a range of experimental conditions and a record is kept of which genes have activated under each condition. Genes which activate under similar conditions are said to be interacting [59]. This method is good because it covers a larger range of conditions than the others and was second only to TAP in the percentage of potential pairs considered (at the optimal parameter choices) [168]. However, two or more genes activating at the same time does not necessarily mean that there has been a physical contact between these proteins. Von Mering et al. found a relatively poor accuracy when compared to the other methods they described [168] and the optimal parameter choices for accuracy resulted in the fourth best coverage, of their reference set of interactions, of the six methods.

As described above no single method has a very high accuracy, however, when an interaction is confirmed by two separate methods then accuracy is improved and there is a further improvement when an interaction is confirmed by three separate methods according to Von Mering et al. [168]. For the analysis of yeast in this thesis we use the data set compiled by Bu et al. [25] which consisted of the high and medium confidence interactions identified by Von Mering et al. [168].

Due to the requirement to consider only interactions which were identified by two or three different methods an analysis using a full simplicial complexes approach is not viable because TAP, Yeast Two–Hybrid and Synthetic Lethality are pairwise measures. As a result the only confirmed three or more way interactions that we could include would need to be determined by all three of HMS-PCI, In-Silico detections and Correlated mRNA expression which makes it far less likely that a three way interaction would be picked up. However, by using a clique complex approach, where any set of three nodes which are all adjacent to each other in the model is considered a 2-simplex, we have shown that it is possible to make inferences about nature of the interactions between the proteins through the simplicial centralities which could not be accomplished without considering this higher topological space. This analysis also provides an example of what could be done in future if more accurate data for the higher order interactions becomes available.

The relationship between protein–protein interaction networks and whether or not a protein was essential was first noticed by Jeong et al. in 2001 [73] who connected the degree centrality of a node to the essentiality of the protein represented by that node. They found that higher degree correlated with higher essentiality. This work was expanded on by Estrada in 2006 who demonstrated that the subgraph and eigenvector centralities outperform the degree centrality when it came to identifying essential proteins particularly where the top 1%, 3% or 5% of nodes are considered [42]. We expand upon Estrada's work by using the simplicial centralities developed in the previous chapter. We demonstrate that, on the Yeast network, by considering the triangular simplicial subgraph centrality it is possible to increase the number of essential proteins identified in the top 1% of the ranking from 50% to 78% and in the top 3% from 50% to 62% compared to the centralities at the node levels. We also note increases from using simplicial degree when compared to degree and simplicial closeness when compared to closeness.

Many other approaches to identifying essential proteins have been employed in recent history which often focus on using edge-based measures in comparison to the traditional node-based ones. For example, the sum of the edge clustering coefficients of a given node was used by Wang et al. [169] to increase the number of identified nodes as a percentage of the top 5%, 10%, 15% and 20% of nodes in the ranking and they tended to identify different essential proteins to the ones found by the node-based measures much like the work in this chapter. The work by Wang et al. also highlighted some difficulties in the data collection of what we have already described as an area of study prone to false positives. They ran the same experiment on three different yeast PPI networks which they had obtained from different sources [169]. Their centrality measure which was based on the edge clustering coefficient was consistently the best performer across the different networks but the performance of the other centrality measures they tested varied across the different networks [169]. For example betweenness centrality was the second best predictor of essential proteins and the subgraph centrality was the worst on one network but these positions were reversed for a different network which represented the protein interactions on the same organism. This lack of a best performing centrality

measure is important because if the betweenness centrality is the strongest centrality measure at detecting essential proteins it would suggest that essential proteins are ones which are perform functions with proteins from many different groups and that there are very few interactions between the members of those groups. However, if the subgraph centrality is the strongest centrality measure then it would suggest that essential proteins are ones which perform functions with many proteins which perform functions with each other. Thus, the choice of a different network representation of the PPI network of the same organism could change the inferences which are made about the proteins in the cell.

Li et al. [93] expanded on the work of Wang et al. by combining a modified edge clustering coefficient with a Pearson correlation coefficient of the two proteins being considered based on the level of expression of their genes which represented a further improvement but was only tested on the DIP database. Meanwhile, Jiang et al. [75] did a further expansion using a weighted sum of five different measures which outperformed the Li et al. method on the datasets they used. These methods draw on more data than is available in just the interaction networks which have been constructed based on PPIs. It demonstrates that the science of studying these interactions has gone beyond the pure network-theoretic approaches used 20 years ago.

The object of the work in this thesis is to demonstrate that it is possible to make inferences about how the proteins interact to perform functions in a cell by considering interactions which happen at higher topological levels than the interactions considered at the network level, which suggests that the same may be true of other phenomena modelled by networks. In future it may be a good idea to combine the values derived from the simplicial centralities with other information to detect essential proteins rather than the combination with the node level centralities which has been done to date.

Since the publication of [50] further work has been done in detecting essential proteins including the work of Klimm, Deane and Reinert in 2021 who took the hypergraphs approach and showed that the node degree of a protein–protein interaction hypergraph could be used to detect essential proteins [87]. In 2024 Lawson, Donovan and Lefevre [92] leveraged the eigenvector centrality, that Tudisco and Higham had

introduced [161], to detect both essential proteins and essential complexes, which were hyperedges which consisted of at least 60% essential proteins. The work of Lawson, Donovan and Lefevre was motivated by the idea that proteins which were essential may well be essential because they are involved in essential interactions within the cell [92]. Other recent advances in identification of essential proteins using PPI networks include: the use, by Rout et al., of betweenness centrality with only shortest paths between proteins which have been associated with forms of cancer considered [132]; the application of a neural network to combine the network centrality measures with gene expression data by Chen et al. [29] and; Ye et al. who also combined network centrality measures with gene expression data but employed an evolutionary community discovery algorithm to do so [171].

Here we study 10 protein–protein interaction (PPI) networks. In these networks nodes represent proteins and undirected links represent the interaction between two proteins determined experimentally via the methods described earlier (mostly Yeast Two–Hybrid and TAP). The networks studied correspond to the following organisms: *D. melanogaster* (fruit fly) [65], Kaposi sarcoma herpes virus (KSHV) [163], *P. falciparum* (malaria parasite) [90], varicella zoster virus (VZV) [163], human [133], *S. cereviciae* (yeast) [25], *A. fulgidus* [105], *H. pylori* [94, 123], *E. coli [26]* and *B. subtilus* [113]. We study only the largest (main) connected components of each of these networks, which range from 50 to 3039 proteins. We transform these networks into their clique simplicial complexes consisting of edges, triangles and tetrahedrons. As described above, it would be preferable to consider three and four way interactions separately from trios of pairwise interactions but it is not possible with the data currently available. The number of simplices and interactions at the nodes, edges and triangle level are given in Table 4.1. We stop at the triangle level because consideration of the tetrahedron level would require knowledge of which cliques of 5 nodes exist and which ones are simplex adjacent to each other which is more computationally intensive than for cliques of 4 nodes. Notice that the number of simplices at the edge level is the same as the number of interactions at the node level.

	nodes		edges	triangles	
species	simplices	interact.	interact.	simplices	interact.
A. fulgidus	32	37	101	1	0
KSHV	50	114	606	34	82
VZV	53	148	1156	104	343
B. subtilus	84	98	463	4	1
P. falciparum	229	604	4599	201	401
E. coli	230	695	7803	478	2425
H. pylory	710	1396	14736	76	79
S. cereviciae	2224	6609	99882	3530	15004
human	2783	6007	85617	1047	2170
D. melanogaster	3039	3687	11369	163	113

Table 4.1: Number of simplices and their interactions at the nodes, edges and triangles levels for the 10 PPI networks studied. Notice that the number of simplices at the edge level is the same as the number of interactions at the node level. This table previously appeared in [50].

4.2 Degree Distributions

We discussed the importance of hub nodes in Section 2.2. In the case of PPI networks hubs are expected to play a fundamental role in the cell and their knockout is expected to cause wide spread cellular damage which is the main finding of the centrality-lethality paradigm established by Jeong et al. [73]. Many PPI networks were characterized as scale free during a fad of classifying networks as such at the beginning of this century, however, since then some authors have found that almost none of the PPI networks previously claimed to have scale-free structures actually do [150]. Blumenthal et al. report that the previous findings may be have been due to study bias and false positives in the interaction datasets [19]. The degree distribution has retained its importance in the study of PPI networks over the years. Khojasteh, Khanteymoori and Olyaee examined degree distributions of PPI networks of SARS-CoV-2 and H1N1 flu viruses in 2022 [83] and Ramos, Ferreira and Simao compared the degree distributions of four different PPI networks for humans in 2024 [128]. The main message of these experiments is that most PPI networks have heavy-tailed degree distributions, such as power-law, lognormal, Burr, logGamma, Pareto, etc.

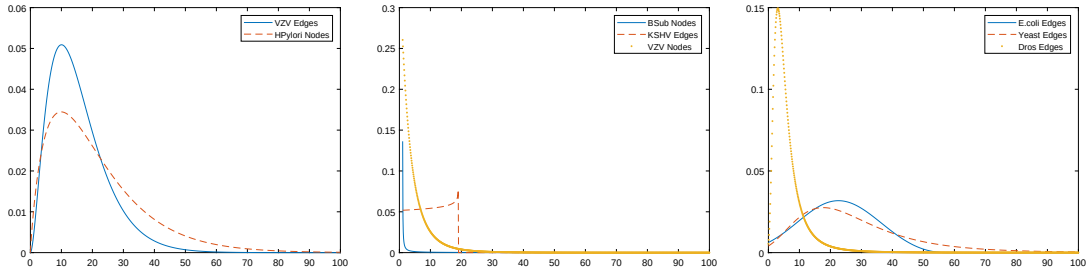
We consider the probability degree functions (PDF), $p(\delta_k)$ vs. δ_k, for 10 PPI net-

species	nodes	edges	triangles
A. fulgidus	NA	$gP(-1.0, 11.5, 0)$	NA
KSHV	NA	$gP(-1.1, 19.3, 1)$	$gP(-1.1, 8.4, 1)$*
VZV	$gP(0.2, 3.8, 1)$*	$gP(-1.0, 31.8, 1)$/ $\Gamma(2.8, 5.6)$**	NA
B. subtilus	$gP(15.9, 4 \times 10^{-15}, 1)$	NA	NA
P. falciparum	NA	$\Gamma(2.7, 5.7)$	$gP(-0.5, 5.8, 0)$
E. coli	$gP(26.9, 8 \times 10^{-15}, 1)$	$GEV(-0.3, 12.2, 18.3)$*	$gP(-0.7, 16.3, 0)$
H. pylory	$gP(23.9, 6 \times 10^{-15}, 1)$	$\Gamma(1.9, 11.3)$	NA
S. cereviciae	$gP(26.2, 8 \times 10^{-15}, 1)$	$GEV(0.1, 13.4, 19.0)$	$gP(-0.4, 11.7, 0)$
human	$gP(25.1, 7 \times 10^{-15}, 1)$	NA	NA
D. melanogaster	$gP(21.8, 6 \times 10^{-15}, 1)$	$GEV(0.3, 2.6, 3.7)$	$gP(23.7, 6 \times 10^{-15}, 0)$

Table 4.2: Degree distributions of the nodes, edges and triangles in the simplicial complexes representing 10 PPI networks studied here (see text for selection criteria). A Generalised Pareto distribution with shape parameter k, scale parameter σ, and threshold parameter θ is represented by $gP(k, \sigma, \theta)$. A Generalised Extreme Value distribution with shape parameter k, scale parameter σ, and location parameter μ is represented by $GEV(k, \sigma, \mu)$. A Gamma distribution with shape parameter a, and scale parameter b is represented by $\Gamma(a, b)$. Not available (NA) distributions are reported when the data was too scarce for a statistically significant fit of the distributions or the statistical criteria used were unable to decide between two or more distributions. *BIC criterion indicates only a strong differentiation with the second best distribution. **BIC criterion indicates only a positive differentiation with the second best distribution (see Figure 4.1). A version of this table previously appeared in [50].

works at the three different levels studied in this work, i.e., nodes, edges, and triangles. For each of the PDFs we fit the data to each of the following distributions: Beta, Binomial, Birnbaum-Saunders, Burr, Exponential, Extreme Value, Gamma, Generalized Extreme Value (GEV), Generalized Pareto (gen-Pareto), Half-Normal, Inverse Gaussian, Kernel, Logistic, Loglogistic, Lognormal, Nakagami, Negative Binomial, Normal, Poisson, Rayleigh, Rician, Stable, t Location-Scale, and Weibull. The best fit was determined on the basis of the following statistical parameters: Akaike information criterion (AIC) [88, 152] and the Bayesian information criterion (BIC) [88] which were discussed in Section 2.7.1. The distributions were ranked in increasing order of their AIC with ties broken using the difference in the BIC for the distributions. In Table 4.2 we show the best distribution fitted for each of the datasets studied.

The most interesting observation from the results shown in Table 4.2 is that while



(a) (b) (c)

Figure 4.1: Examples of the Gamma Distribution (a) (blue solid line is the edge degree distribution for VZV, orange dashed line is the edge degree distribution of H. pylori), Generalised Pareto Distribution (b) (blue solid line is the node degree distribution for B. subtilis, orange dashed line is the edge degree distribution of KSHV, yellow dotted line is the node degree distribution for VZV) and Generalised Extreme Value Distribution (c) (blue solid line is the edge degree distribution for E. coli, orange dashed line is the edge degree distribution of S. cereviciae, yellow dotted line is the edge degree distribution for D. melanogaster).

the degree distributions obtained for the node level of the PPI simplicial complexes are heavy-tailed this pattern is not necessarily repeated for the 1-simplices and 2-simplices. At the node level, the 7 distributions that were statistically significant correspond to a generalized-Pareto distribution with a positive shape parameter, where the probability of finding nodes of a given degree decays as a power-law of the corresponding degree (see the blue solid line and yellow dotted line of Figure 4.1 (b)). For the other three datasets the statistical criteria used were not able to distinguish between the first few distributions.

At the 1-simplex level only E. coli, S. cereviciae and D. melanogaster display a Generalised Extreme Value distribution (see Figure 4.1 (c)) which is heavy tailed but does not feature a high proportion of low degree simplices as was the case for the node degree distributions. Instead we see a varied range of degrees centred around 19 for E. coli and S. cereviciae and around 4 for D. melanogaster before the start of the long tail. This finding suggests that while there are pairs of interacting proteins which have very high degree and are likely to be very important to the functioning of the cell there are also a lot of pairs of interacting proteins which have a medium degree and comparatively few of low degree. The rest of the organisms display a similar style of

degree distribution for the edges except without the large tail. Neither the Gamma distribution (see Figure 4.1 (a)) nor the generalized-Pareto distribution with a negative shape parameter (see orange dashed line of Figure 4.1 (b)) have heavy tails. Both of these distributions also feature a large number of medium degree simplices. The large number of medium degree 1-simplices suggests that the loss of the interaction between these proteins should not be destructive to the functioning of the cell.

At the 2-simplex level all of the PPI simplicial complexes for which it was possible to determine the best distribution display generalized-Pareto distributions. The shape parameter of these distributions was negative for all of them with the exception of D. melanogaster which suggests that there is a slight increase in the number of high degree 2-simplices compared to low degree ones with many medium degree ones in between and that the loss of the interaction between a trio of nodes should not be very destructive to the cell. However, it is often the case that many of these high degree interactions have proteins in common and the loss of the individual proteins which feature in many high degree interactions can be very destructive for the cell as we demonstrate in Section 4.4.

4.3 Comparison of Centralities at Different Levels

Simplicial centrality measures are all designed to identify the "most important" simplices in a simplicial complex, at different levels, according to certain topological features of the complex, such as nearest-neighbour connectivity (degrees), proximity of other simplices (closeness) and participation of a simplex in small sub-complexes with other simplices (subgraph centralities). It should be expected that there is some correlation between the centralities inside each level of analysis. That is, we would expect that node degree is somehow correlated to node closeness or node subgraph centrality for a given PPI network. We employ Spearman's rank correlation coefficient which was introduced in Section 2.7.2 to assess the correlation between the different centrality measures. We choose to use Spearman's correlation coefficient because we expect that there will be differences in the distribution of the results for each of the centrality measures chosen. For example the subgraph centrality is the result of an exponential matrix

function while closeness and degree are not. Additionally, with centrality measures it is, generally, the ranking that is important rather than their magnitude of the result.

In the yeast PPI the node centralities (degree, closeness and subgraph centrality) have an average rank correlation coefficient $\langle rs_{n,n} \rangle \approx 0.828$, with the largest rank correlation coefficient rs recorded between the closeness and the subgraph centralities ($rs(SC, CC) \approx 0.924$). At the edge level this average rank correlation is $\langle rs_{e,e} \rangle \approx 0.827$ and at the triangle level it rises up to $\langle rs_{t,t} \rangle \approx 0.970$.

We need to check that the simplicial centralities are actually providing new information because it could be the case that the highest ranked edges and triangles consist of the highest ranked nodes which would mean that no new insights can be gained by considering the centralities at the higher levels. We take each node and calculate the average centrality of each edge it is involved in. We then rank the nodes based on the average centrality of the edges which allows us to compare the rankings that the edges give to the node rankings through the Spearman rank correlation coefficient. We also perform this aggregation for the triangles so that the simplicial centrality rankings of the 2-simplices can be compared to the rankings for the 1-simplices and 0-simplices. The average rank correlation coefficient between the node and edge centralities is just $\langle rs_{n,e} \rangle \approx 0.609$, and between the node and triangle centralities it is $\langle rs_{n,t} \rangle \approx 0.587$. Finally, the average rank correlation between the edge and triangle levels is $\langle rs_{e,t} \rangle \approx 0.228$. The correlation between the inter-level centralities (see Table 4.3) is lower than between the different centrality measures at the same level which suggests that it is not the case that the triangle centrality rankings consist simply of 2-simplices made up of the highest ranked nodes.

It is also important to consider that none of the correlations are negative. This fact implies that none of the centralities fundamentally disagree with each other. It is not the case that a centrality at one level is telling us that one set of nodes is not important and another set of nodes is but a centrality at a different level is telling us the exact opposite. It is more likely that a centrality at one level is telling us that one set of nodes is important while a centrality at a different level is telling us the same thing but the order of importance is shuffled between the two centralities. This hypothesis is backed

	nodes		edges			triangles		
	SC	CC	DC	SC	CC	DC	SC	CC
Node Degree	0.7602	0.7984	0.3292	0.3856	0.3020	0.6586	0.6951	0.6812
Node Subgraph		0.9245	0.7376	0.6981	0.7471	0.5562	0.5780	0.5990
Node Closeness			0.7347	0.7710	0.7805	0.4795	0.5103	0.5260
Edge Degree				0.7617	0.9025	0.2744	0.2794	0.2928
Edge Subgraph					0.8180	0.2042	0.2207	0.2456
Edge Closeness						0.1558	0.1691	0.2076
Triangle Degree							0.9725	0.9772
Triangle Subgraph								0.9589

Table 4.3: Spearman's rank correlation coefficients between the rankings of three centralities of the $0, 1$ and 2-simplices in the yeast PPI simplicial complex. This table previously appeared in [50].

up when we consider the triangle and node degrees. Of the 100 most central nodes according to these centralities 24 coincide. When we consider the top 300 there are 111 (37%) proteins in common and looking at the top 500 the two centralities identify 268 (53.6%) common proteins. This change in the percentage of common proteins at the top of the rankings between the nodes and triangles may explain the difference between the centralities capabilities in the detection of essential proteins in the next section when a small percentage of the top proteins are considered compared to the similarity when a larger percentage is considered.

We can expand this study of the average rank correlation coefficient to all the PPI simplicial complexes considered in this work. In Table 4.4 we give the average Spearman rank correlation coefficients for all of the PPI networks studied. We can see that the intra-dimensional correlations between the centralities considered is relatively high which follows our expectation that different centralities identify essentially the same groups of proteins at each corresponding level. The highest correlations are observed for the triangle level, which is mainly due to the high correlation between the triangle degree and closeness centralities. This high correlation could be a consequence of the fact that most of the high degree triangles are clumped with many other triangles (see next section for the case of yeast). These high-degree triangles are close to each other, giving a high triangle closeness. Finally, we also observe lower rank correlation between the different pairs of levels considered for the 10 PPI networks analysed. The

	$\langle rs_{n,n} \rangle$	$\langle rs_{e,e} \rangle$	$\langle rs_{t,t} \rangle$	$\langle rs_{n,e} \rangle$	$\langle rs_{n,t} \rangle$	$\langle rs_{e,t} \rangle$
A. fulgidus	0.822	0.844	NA	0.399	NA	NA
KSHV	0.912	0.778	0.993	0.632	0.666	0.493
VZV	0.873	0.783	0.899	0.176	0.751	0.186
B. subtilus	0.749	0.865	0.792	0.407	0.294	-0.030
P. falciparum	0.842	0.826	0.957	0.608	0.655	0.403
E. coli	0.842	0.915	0.959	0.714	0.699	0.483
H. pylory	0.847	0.740	0.932	0.608	0.458	0.254
S. cereviciae	0.828	0.827	0.970	0.609	0.587	0.228
human	0.732	0.818	0.929	0.641	0.505	0.235
D. melanogaster	0.661	0.703	0.795	0.568	0.303	0.188

Table 4.4: Intra- ($\langle rs_{n,n} \rangle$, $\langle rs_{e,e} \rangle$ and $\langle rs_{t,t} \rangle$) and inter-level ($\langle rs_{n,e} \rangle$, $\langle rs_{n,t} \rangle$ and $\langle rs_{e,t} \rangle$) average Spearman's rank correlation coefficients between the rankings of three centralities of the $0, 1$ and 2-simplices in the 10 PPI simplicial complexes studied. See text for notation and explanations. This table previously appeared in [50].

slightly negative average obtained for $\langle rs_{e,t} \rangle$ in *B. subtilus* can be considered more as a lack of correlation than a negative correlation between the simplicial levels.

4.4 Identification of Essential Proteins

4.4.1 Methodology

An essential protein is one that when knocked out results in the death of the cell. The identification of such proteins has become one of the main paradigms of the study of centrality measures in PPI networks [29, 42, 69, 73, 75, 87, 92, 93, 132, 141, 169, 171, 173]. The reasons for this interest are twofold. On the one hand, it is important to have theoretical tools that allow the identification of proteins that can be drug targets, think of the identification of essential proteins in a pathogenic microorganism. On the other hand, it is one of the scarce examples in which centrality measures can be validated against experimental data. The methodology for essential protein identification that we consider here is adapted from the one developed by Estrada in 2006, and consists of the following steps [42]. First, we transform the PPI network into a clique simplicial complex with cliques of up to 4 nodes turned into simplices which allows us to consider 0-simplex, 1-simplex and 2-simplex centralities. Note that to consider the centralities

at the 3-simplex level then the cliques of 5 nodes would need to be represented in the simplicial complex as 4-simplices because to know whether or not two 3-simplices are simplex adjacency then it is necessary to know if they are both faces of the same 4-simplex. However, knowledge of the existence of $(k+n)$-simplices, where $n \geq 2$ does not affect the simplex adjacency of the k-simplices which means the existence of the 3-simplices in the PPI simplicial complex of yeast is enough to calculate the node, edge and triangle centralities. Secondly, we calculate the corresponding centralities at the three levels for each of the simplices. We then transform edge and triangle centrality into a ranking of the nodes by calculating the average centrality of all the edges and triangles in which a given node is involved, as per the previous section. Using these centralities we rank all of the proteins in the PPI in decreasing order. Finally, we count how many essential proteins are in the top $x\%$ of the ranking and report this number as the percentage of essential proteins identified by the corresponding centrality (see Figure 4.2). An ideal index for essential protein identification is one which ranks all essential proteins at the top of the ranking, such that if we want to select 100 essential proteins we simply pick the top 100 proteins in that ranking. We compare the results of this process with the random selection of proteins. That is, we rank the proteins randomly and count the essential ones which are in the top $x\%$ of the ranking.

4.4.2 Application to Yeast PPI

We now apply the methodology previously described to identify essential proteins in the yeast PPI. There are several datasets of the interactions of proteins in yeast. Here we use the data compiled by Bu et al. [25]. The original data was obtained by Von Mering et al. [168] by assessing a total of 80,000 interactions among 5400 proteins reported previously and assigning each interaction a confidence level. Bu et al. [25] focused on 11,855 interactions between 2617 proteins with high and medium confidence in order to reduce the interference of false positives, from which they reported a network consisting on 2361 nodes and 6646 links:

`http://vlado.fmf.uni-lj.si/pub/networks/data/bio/Yeast/Yeast.htm`.

We transform this interaction map into the clique complex of a network in which

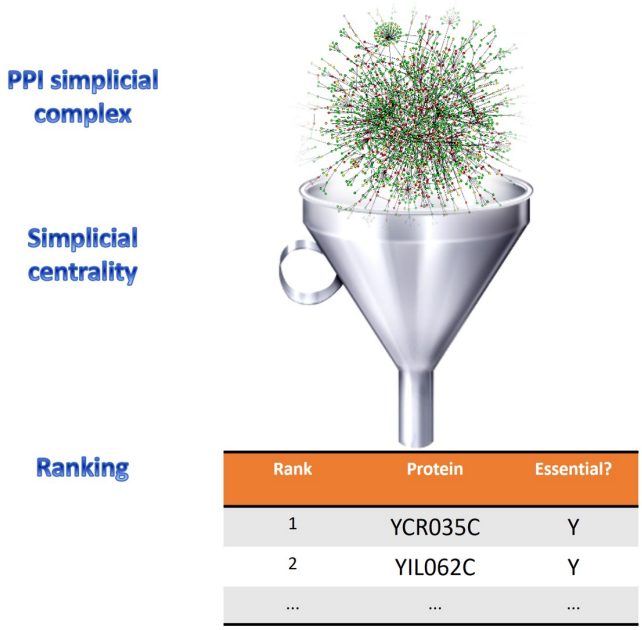


Figure 4.2: Schematic representation of the process of identification of essential proteins using simplicial centralities in a PPI. This figure previously appeared in [50].

proteins are represented as nodes and two nodes are linked by an edge if their corresponding two proteins interact with high or medium confidence. In this section we consider the 0-simplex, 1-simplex and 2-simplex degree, closeness, and subgraph centralities as examples of nearest-neighbour, shortest-path and local neighbourhood centralities respectively.

The first interesting observation obtained from this analysis is that the centralities based on the edge level of the simplicial complex perform very badly at identifying the essential proteins. For instance, for the top 1% of proteins in the ranking, the node and triangle centralities identify more than 45% of essential protein (see detailed analysis below), but the edge centralities do not identify more than 27% (edge degree). In general, none of the edge centralities are able to identify more than 35% of essential proteins at any percentage of top proteins selected. This result contrasts with the rankings obtained by using node and triangle centralities. For instance, for the closeness centrality at both node and triangle level, the number of essential proteins identified is always larger than 37%. As can be seen in Figure 4.3(b) the triangle closeness centrality

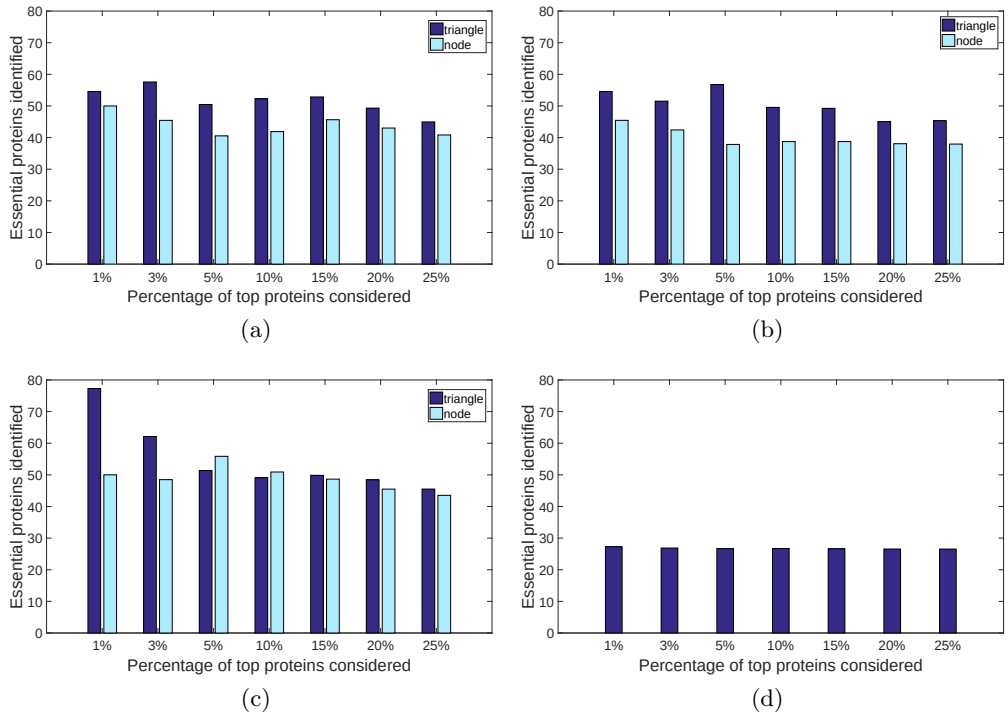


Figure 4.3: Percentage of essential proteins identified using simplicial degree (a), closeness (b), simplicial subgraph centrality (c) and random selection (d) based on the 0-simplices and 2-simplices of the simplicial complex for yeast PPI. For the case of the random selection, as the edge and triangle information is reduced to values for the nodes, only the selection of essential proteins based on random ranking of the nodes is required. This figure previously appeared in [50].

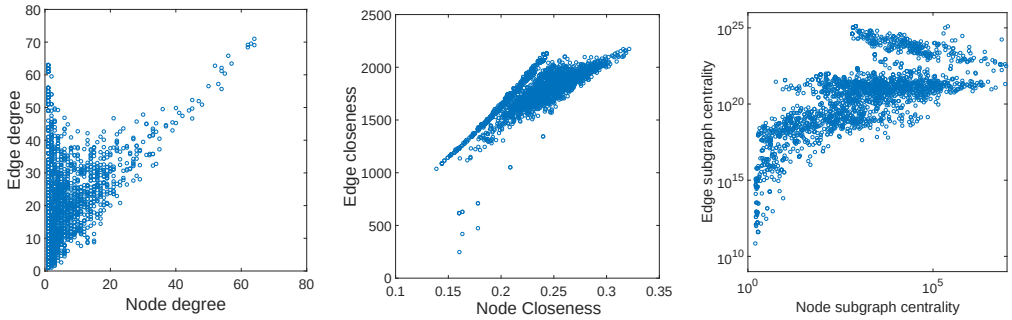
outperforms the node version for all of the percentages of proteins considered. In the top 10% and 15% of proteins ranked by the triangle closeness an extra 10% of the proteins identified are essential compared to the equivalent criteria for nodes. These differences represent up to 40 additional essential proteins identified by the triangle centrality.

The largest percentages of essential proteins identified are obtained by the subgraph centralities. In particular, for 1% and 3%, the triangle subgraph centrality greatly outperforms the node one. For the top 1% of proteins identified by the triangle subgraph centrality an extra 20% of the proteins identified are essential compared to its node analogue and for the top 3% it outperforms the node centrality by 14%. However, for top percentages of rankings over 5%, the node and triangle subgraph centrality do not

show large differences in the identification of essential proteins. Around 50% of the proteins identified are essential for both rankings.

The fact that the edge centralities do not describe the essentiality of proteins in the yeast PPI simplicial complex merits explanation. This observation clearly indicates that increasing the complexity of the representation of a phenomenon does not imply that the amount of information which can be extracted from that system also increases. For simplicity, we provide an explanation for why the degree of the edges does not detect essential proteins but the explanation is also valid for the other centralities studied here. Let us recall from Theorem 3.15 that the edge degree is given by $\delta_1(\sigma) = \delta_0(i) + \delta_0(j) - (2 + 2T)$, where i and j are the nodes forming the edge σ and T is the number of triangles that the edge is a face of. Notice that in a graph theoretical framework the edge degree is simply defined as $\delta_0(i) + \delta_0(j) - 2$. The important thing here is that the edge degree in the simplicial complex depends on the degree of the nodes forming an edge and the number of triangles it participates in. In Figure 4.4 we show the scatter-plots of the edge centrality indices versus their node analogues. As can be seen in all cases the correlation is positive and for the cases of the degree and closeness the correlation between the two centralities is relatively good. We now analyse the causes of the differences between the node and edge centralities and how they influence the edge centrality's inability to identify essential proteins in the yeast PPI.

Suppose that the number of triangles that an edge is a face of is relatively small, such that the degree of the edge is mainly dependent on the degree of the nodes forming that edge. Then, it is possible to have two different edges with exactly the same edge degree which differ significantly in the degree of the nodes forming the edges. That is, we can have an edge formed by two nodes of mid-degree (MD), e.g., MD-MD, and another formed by a high-degree (HD) and a low-degree (LD) node. It is not difficult to find many of these examples in the yeast PPI. For instance, the edges YMR125C-YOL139C and YDR386W-YOL139C are formed by nodes of degrees 39-36 and 31-36, respectively. That is, these two edges are of the MD-MD type. On the other hand, the edges YPR110C-YLR086W and YPR110C-YGL016W are formed both by nodes of

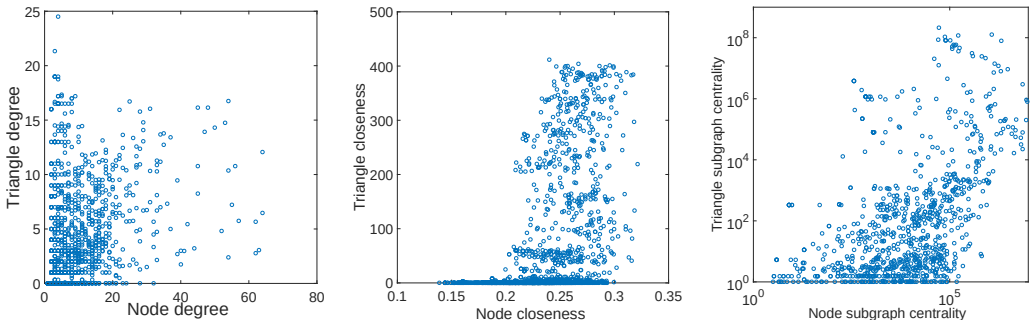


(a) (b) (c)

Figure 4.4: Scatter plots of the edge centralities, degree (a), closeness (b) and subgraph centrality (c), versus the node analogues of the same centralities. Notice that the subgraph centrality plot is in log-log scale and that harmonic closeness was used for the edge closeness centrality due to the Yeast PPI simplicial complex not being 1-connected. This figure previously appeared in [50].

degrees 64-3, which clearly means that they are of the HD-LD type. It is well-known that high-degree nodes are more likely to represent essential proteins (see Figure 4.3(a)). Thus, it is more probable that an HD-HD edge contains an essential protein than an MD-MD one. Indeed, neither of the proteins in the previous example in MD-MD are essential, but the protein YPR110C in the HD-LD edges is. The situation is even worse when the nodes involved in a given edge participate in a large number of triangles. In this case, twice the number of triangles is subtracted from the degree of the two nodes the simplicial degree for the 1-simplices. Thus, if an edge is involved in a large number of triangles, its edge degree will be relatively small due to the fact that it is penalized for each triangle it is a part of. Thus HD-HD edges are unlikely to have a high edge centrality because two HD nodes which are adjacent are likely to have many neighbours in common. The existence of nodes having low degree but displaying either very low or very high edge degree is easy to understand. In edges of the HD-LD type, there is always a node with low degree which displays very large edge degree due to the influence of the HD node. In those edges where an LD node is connected to another LD node, both the node and the edge degree are low. These factors explain the edge centralities failure to predict essential proteins in the yeast PPI compared to the node and triangle centralities.

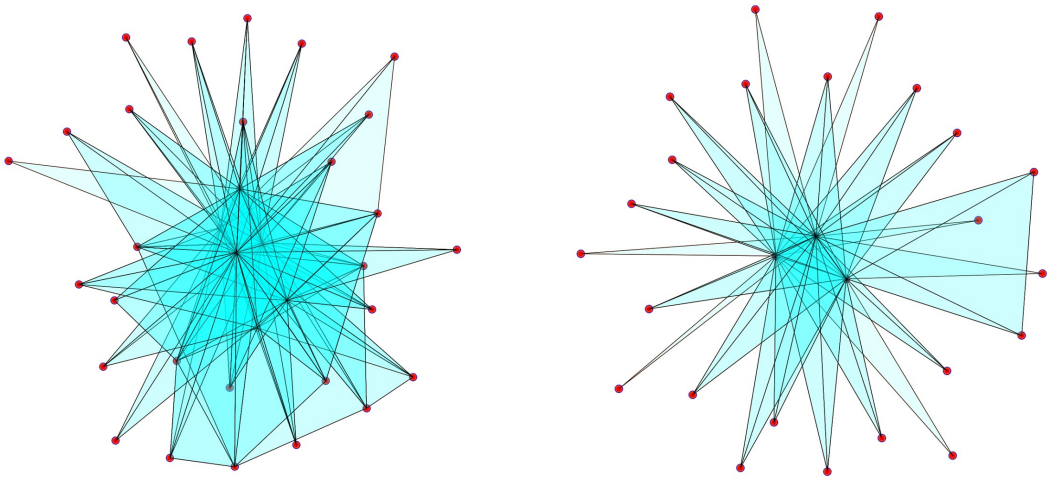
Now we move to the analysis of the triangle centrality indices. Because the triangle



(a) (b) (c)

Figure 4.5: Scatter plots of the triangle centralities, degree (a), closeness (b) and subgraph centrality (c), versus the node analogues of the same centralities. Notice that the subgraph centrality plot is in log-log scale and that harmonic closeness was used for the triangle closeness centrality due to the Yeast PPI simplicial complex not being 2–connected. This figure previously appeared in [50].

centralities outperform the node ones in identifying essential proteins, our main goal here is to identify a structural pattern which contributes to the triangle centralities and does not contribute to the node ones. In order to perform our analysis we again consider the degree centralities for the sake of simplicity. We are only interested in the structural information which is useful for the identification of essential proteins. The structural pattern that we identify here consists of a node A which is the vertex of a relatively small number of triangles, such that its node degree is small. Suppose for instance that A is connected to the nodes B, C, and D forming the triangles ABC and ACD. Obviously the node degree of A is only 3. Now, let us consider that BC is an edge of a large number of triangles, and that CD has a similar property which means that the triangles ABC and ACD have large triangle degree and consequently the node A is very central according to this index. A node is highly triangle-central if the edges of the 2-simplices it is a member of are also members of a large number of other 2-simplices. We provide two complexes displaying exactly this structural pattern as examples. The first is formed by the protein YDL148C, which is connected to 4 other proteins, namely YGR090W, YBR247C, YCL059C and YCR057C. These proteins form 5 triangles in which YDL148C is a vertex. Then, obviously, the protein YDL148C is not very central according to this nearest-neighbour structure, i.e., its node degree is



(a) (b)

Figure 4.6: Illustration of the two simplicial complexes in formed by the proteins YDL148C (a) and YMR112C (b). This figure previously appeared in [50].

only 4 and it participates in only 5 triangles. However, the edges of these 5 triangles participate in a total of 88 other triangles. That is, the edge YGR090W–YBR247C takes place in 11 other triangles, YGR090W–YCR057C in 22, YBR247C–YCL059C in 14, YCL059C–YCR057C in 25, and YBR247C–YCR057C in 16. The protein YDL148C is then very central according to the triangle centrality and it is essential. Another example is provided by the protein YMR112C, which is also essential and is connected only to YDL005C, YOL135C, YBR253W and YJR068W. It forms only 4 triangles, but the edges of these triangles form 14, 15, 17 and 19 other triangles, respectively. Thus, the protein YMR112C which is not central according to node centrality is one of the most central nodes when the 2-simplex centrality indices are reduced to the node level.

It should also be noted that there is structural information contained in the node centralities which is not accounted for by the triangle ones. As we have seen before there are proteins with high node centrality and low triangle centrality. However, the number of structural patterns contributing to this situation is wider and ranges from the simplest case where a protein interacts with a large number of other proteins which do not interact with each other, to the case where a central protein forms a wheel-like structure. In the first case obviously the protein has a high degree but its triangle

degree is zero. In the second case a central node is connected to every node of a cycle having $n-1$ nodes, the central node has degree $n-1$ but every triangle has degree only two. The important message of this section is that the triangle centrality includes some structural information which is relevant for understanding biological processes such as the essentiality of proteins in the yeast PPI.

Chapter 5

Random Geometric Simplicial Complexes

Parts of this chapter have been presented at the University of Strathclyde Applied Analysis Seminar and at the International Congress on Industrial and Applied Mathematics (ICIAM) 2019.

5.1 Literature Review

We now consider random geometric graphs (RGGs) which were first introduced, under the guise of random plane networks, by Gilbert in 1961 [63] as a foil to the Erdős-Rényi Random Graphs introduced a year earlier [41]. Gilbert was motivated to introduce them because in an Erdős-Rényi random graph any two nodes are connected with an equal probability. This property is acceptable in the null model when the graph system being modelled is not embedded in a topological space but becomes less relevant when it is. The example Gilbert gives is of short range radio stations which can communicate only when they are within a certain distance of each other. If the same system were modelled by the Erdős-Rényi random graph then each radio station would be equally likely to be able to communicate with any other radio station regardless of location. Random Geometric Graphs were later extended to more than two dimensions [119].

Another key application of RGGs is in wireless sensor networks (WSNs) [1, 10, 56,

61,82,106,121,127,155] which are networks of sensors which can detect changes in their environments. They are usually limited in their detection capacities to a given area, examples include the underwater wireless sensor networks which are used to detect tsunamis [10] or the WSNs used on nuclear sites to detect leaks [82]. They also have applications to monitoring hillsides for landslides [127] and habitats to search for particular species [121]. Recently WSNs have become an important component of "Internet of Things" systems where the environment in a home, hospital [61], piece of industrial machinery [56] or smart agricultural system [106] is monitored using wireless sensors and then decisions about changes to the environment are made by a centralised base station [1] indeed Tamene et al. [155] view the Internet of Things to be the natural successor to traditional WSNs. The recent proliferation of applications for WSNs has driven a focus on energy efficiency [164] and the related question of their security in order to meet requirements to protect data [60]. There are two ways to model WSNs: in the case that the sensors are able to communicate with each other then if two sensors were positioned closely enough together such that communication was possible then they could be connected by an edge; similarly if the regions that two sensors cover overlap that could be another condition where we would connect two nodes in the random geometric graph.

It is possible for three or more sensors to be in communication with each other and it is also possible that three or more sensors would cover a common area. As a result RGGs have been expanded to consider interactions between more than two nodes. A Random Geometric Hypergraph consists of a set of nodes spread on the unit d-dimensional cube $[0,1]^d$ and a connection radius r and then x nodes are members of the same hyperedge if there is a common intersection of the balls of radius r centred at those nodes. However as Turnbull et al. [162] and Kahle [78] point out under this definition the balls of radius r centred at any subset of these x nodes would also have common intersection and so any subset of the hyperedge containing these x nodes would also be a hyperedge and so we naturally arrive at the simplicial complex structure. However, other notions of random geometric hypergraphs have been proposed. One example would be De Kergorlay and Higham's construction [33] where there are various hypergraph centres and each node

would join the hypergraph centres to which it was sufficiently close. In their example they considered that the hypergraph centres were meeting places like shops, gyms or churches and individual people would go to the ones closest to their homes. It is easy to see how this model could be extended to larger or more complex internet of things set ups with multiple base stations. Although this new construction is interesting we do not consider such complicated cases here.

There are two ways to form simplicial complexes from random geometric graphs. The first way is known as the Čech complex [62] and the second is the Vietoris–Rips complex [62, 167]. Kahle [77] has studied the properties of these random geometric complexes when embedded in different dimensional spaces. He calculated the expected Betti numbers of such a complex under different conditions. Betti numbers are particularly relevant to the applications which we mention above and we discussed them in Section 2.6. The Betti numbers of a simplicial complex describe the number of gaps of that dimension which exist within the complex. The WSNs we discuss here are embedded in a 2-dimensional space so it only really makes sense to talk about gaps in two dimensions, usually known as holes. The first Betti number, β_1, of a Random Geometric Simplicial Complex is equal to the number of these two dimensional holes. In wireless sensor networks these holes would represent gaps or dark zones in the coverage of an area. In fact, a lot of work has gone on in the literature into detecting these gaps in coverage using a simplicial complexes approach [34, 107, 126].

One deficiency of limiting random geometric simplicial complexes to a single connection radius when studying wireless sensor networks is that while it is possible to know the number of holes there is no information about their size. For a wireless sensor network it may not be a problem to have many small gaps in the coverage but a single large gap could be very problematic if the job of the sensor network is to prevent natural disasters. One solution to this problem is De Silva and Ghrist's use of the eigenvectors associated with the largest eigenvalues of the Hodge Laplacian at the 1-simplex level which they showed tends to have higher values for edges which were on the boundary of the largest hole [34]. Another solution to this problem is persistent homology with larger holes expected to persist for longer. Chintakunta and Krim leveraged persistent

homology to check for false positives in their coverage hole detection algorithm [30]. Bobrowski, Kahle and Skraba introduced a multiplicative persistence where the death time of a hole was divided by its birth time to create a ratio which allows persistence of holes to be compared across different random geometric simplicial complexes [20]. They also provided some probabilistic proofs about the nature of their multiplicative persistence. Persistent homology has also been used to detect community structures on Random Geometric Graphs among many other graphs by Jenkins [72].

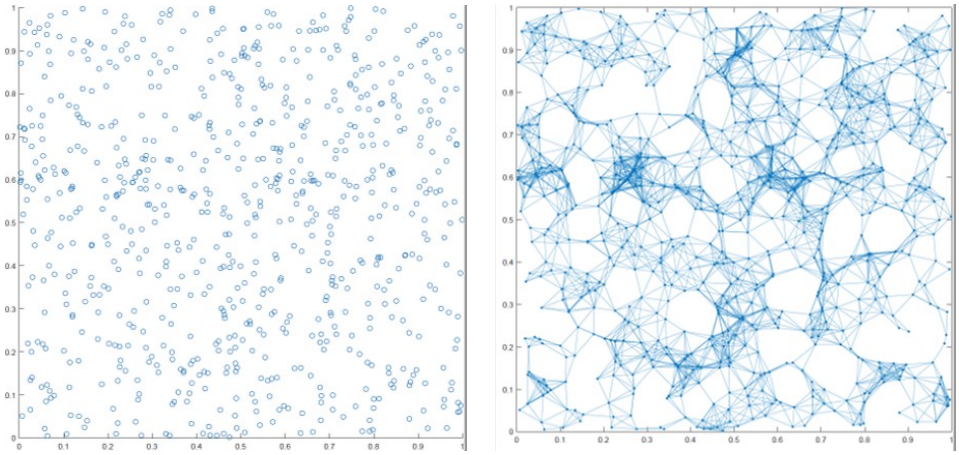
This interest in detecting gaps in coverage is primarily so that it is possible to plug those gaps and maintain communication or coverage across these regions. However, as Kenniche and Ravelomananana [82] note, there are wireless sensor networks where the sensors are fitted with batteries which it may not be possible to recharge or replace. For example, the sensors, in the WSN that Polastre et al. used to monitor the Leach's Storm Petrel on Great Duck Island, failed for many different reasons [121]. Sensors may also fail for other reasons and it may not be possible to replace them quickly, for example a tsunami detection sensor in a remote part of the ocean, or safely, for example during heavy rain on a hillside which is known to be prone to landslides. Given the potentially fatal consequences of sensors failing much research has been done into detecting this situation [102, 109] and tolerating it [27, 172]. Many of the recent advances in failure detection have centred around the use of machine learning algorithms such as the use of extremely randomised trees by Saeed et al. [135] and the application of an adaptive shark smell optimisation algorithm in combination with a modified Elman recurrent neural network by Mahalakshmi et al. [96]. Ghadi et al. provided an overview of the machine learning algorithms in use in the study of WSNs in 2024 [60].

There are two general types of approaches to building tolerance to failures in a WSN: creating fault tolerant node distributions when the WSN is introduced and moving existing nodes or introducing new nodes to the system to act as redundancies in the situation where a node failure would lead to a gap in coverage [172]. In WSNs with mobile nodes algorithms have been introduced to decide where and when to move nodes such as the one developed by Tirandazi et al. which combines a local algorithm and a global algorithm to allow the sensors to make these decisions [157]. Sadeghi Ghahroudi

et al. provided a survey of the available algorithms for node movement in WSNs in 2023 [134].

Introduction of new sensors is of particular importance in the case that the sensors are not mobile and so cannot be programmed to cover gaps in coverage themselves. One strategy for introducing nodes to eliminate holes in the coverage of a WSN is the one proposed by Simionato and Cimino [146]. Their strategy is to have collections of mobile sensors located at points throughout the area being monitored. When a hole is detected these sensors are sent to the boundary of the hole which shrinks the boundary of the hole as more sensors show up to sit on the reduced boundary until the hole is no longer present and any remaining sensors can return to their base [146]. These new sensors would remain in place until the failed sensor can be repaired or replaced at which point the mobile sensors used to cover the gap would become available to deal with new gaps [146].

Dependent on how the sensors were originally placed and the limitations on these placements then some areas of the region being monitored may be more prone to gaps in coverage opening up than others. These areas are ones in which there are multiple nodes whose failure would create a coverage hole which would determine where the optimal places for Simionato and Cimino's collections of mobile sensors would be. It is useful to be able to update this knowledge quickly in the event of failures of other nodes in the system in order to move these collections or know which sensors to prioritise fixing. Persistent homology could provide hints towards which nodes are likely to induce a hole if they were removed. Nodes which were among the last nodes to join the large connected component or nodes that were on the boundary of holes that died at radii just below the connection radius would be good candidates. However, nodes may join smaller connected components before the construction of the large connected components and the identifiers for these smaller connected components would not be unique to these hole-critical nodes. Similarly, a node which is positioned centrally to a set of other nodes which are at a distance of slightly over the connection radius to each other and three quarters to two thirds of the connection radius to the centrally located node would not be on the boundary of a recently deceased hole but would



(a) (b)

Figure 5.1: 1000 nodes scattered on a square of side length 1 (a), The RGG constructed from these nodes with a connection radius of 0.07 (b).

induce a hole if it was removed and be covering a region on its own if a WSN was being modelled. This chapter presents a method to identify nodes which are likely to create a homological hole if they were removed.

5.2 Introduction

We begin by providing an extension of the definition of the Random Geometric Graph to multiple dimensions from Penrose [119].

Definition 5.1. The **random geometric graph (RGG)** is defined by distributing n nodes independently and uniformly in the unit d -dimensional cube $[0,1]^d$. We then say that two nodes are adjacent if their Euclidean distance is less than r which is a parameter known as the **connection radius**.

If we look at Figure 5.1 we can see the process of constructing an RGG in action with 1000 nodes scattered on the unit square on the left and these nodes connected if they are within a distance of 0.07 of each other. We are operating in 2–dimensional space which is the case for the rest of the thesis.

We now extend random geometric graphs to random geometric simplicial complexes using Kahle's definition [77].

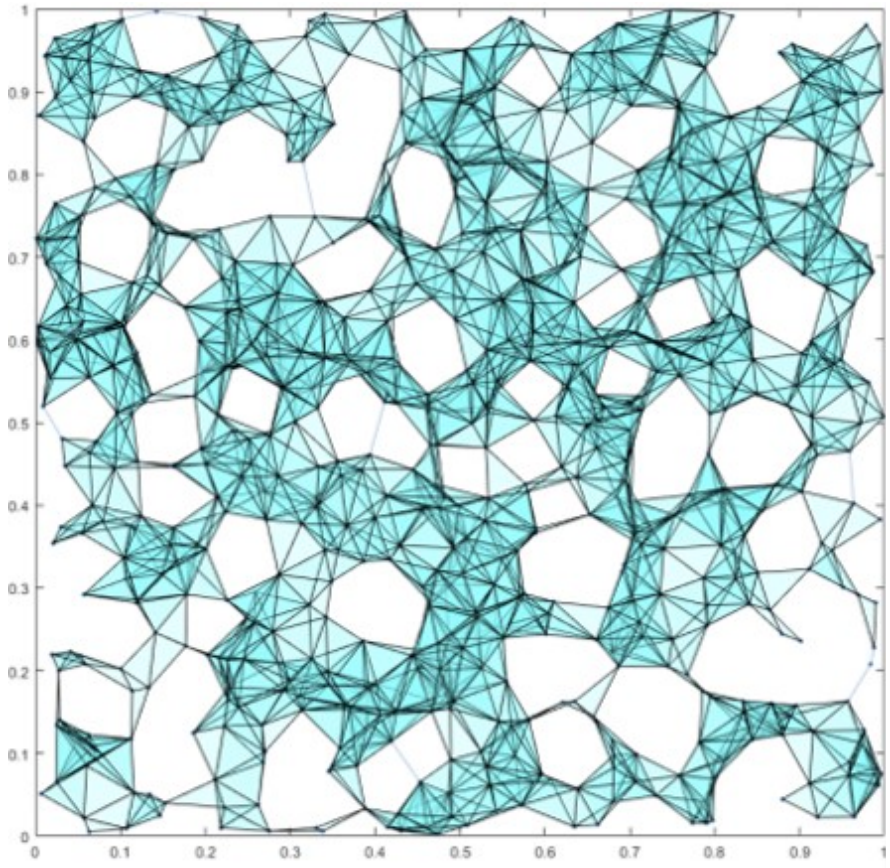


Figure 5.2: An RGSC constructed from the RGG in Figure 5.1 using the Vietoris–Rips complex method.

Definition 5.2. A **random geometric simplicial complex (RGSC)** is a simplicial complex defined using the nodes and edges of an RGG as the 0-simplices and 1-simplices respectively.

We can see the effect of extending the random geometric graph from Figure 5.1 into an RGSC in Figure 5.2. We shall demonstrate how to create holes in RGSCs which correspond to gaps in coverage of a WSN.

There are many different definitions by which we could extend random geometric graphs to random geometric simplicial complexes. The most useful method would be to use the Čech complex [62, p. 30].

Definition 5.3. To construct the **Čech complex**, $\mathscr{C}$, of an RGG we consider the ball of radius $\frac{r}{2}$ around each node i, $B_{\frac{r}{2}}(i)$. Now for every subset of the nodes $X \in V$, then

$X \in \mathscr{C}$ if and only if $\bigcap_{i \in X} B_{\frac{r}{2}}(i) \neq \varnothing$.

Let there be three sensors in a WSN and let there be an edge between each pair of nodes represented which means that if the radius of the coverage area for each sensor is $\frac{r}{2}$ then the coverage areas of each pair of these sensors must overlap. However, there is no guarantee that there is an overlap between the coverage areas of all three sensors, i.e. that the intersection of the coverage areas of all three sensors is non-empty. If it is non-empty then we would represent this situation as a 2-simplex containing the three relevant nodes in the Čech complex and if not then we would not include such a simplex. The absence of such a simplex corresponds to a gap in the coverage in the WSN and the presence of a simplex means that any point in the area defined by the convex hull of the location of the three sensors is covered by at least one sensor in the WSN.

Unfortunately, the Čech Complex is very time consuming to determine because of the need to check the intersection of every subset of the nodes. A faster method is the calculation of the Vietoris–Rips complex [62, p. 28].

Definition 5.4. The **Vietoris–Rips** complex of an RGG is its clique complex.

As before, let there be three sensors in a WSN and let there be an edge between each pair of nodes represented. However, this time because all three nodes are mutually connected they form a triangle and so a 2-simplex containing these nodes is automatically included in the Vietoris–Rips complex.

We can see that in terms of providing an accurate representation of the coverage of a WSN then the Vietoris–Rips complex is not as accurate as its Čech equivalent because the presence of a simplex in the Vietoris–Rips complex does not guarantee that every point in the convex hull of the sensors which define the simplex is covered unlike the Čech case. However, Theorem 2.5 of "Coverage in Sensor Networks Via Persistent Homology" by De Silva and Ghrist [34] states that it is possible to squeeze a Čech complex between two Vietoris–Rips complexes of different sizes. We state the theorem below for completeness.

Theorem 5.5. *Let $\mathscr{X}$ be a set of points in $\mathbb{R}^d$, $\mathscr{C}_r(\mathscr{X})$ be the Čech complex of these*

points when we consider balls of radius $\frac{r}{2}$ *and* $\mathscr{V}_r$ *be the Vietoris–Rips complex of the Random Geometric Graph formed by considering the nodes to be the set* $\mathscr{X}$ *and connecting nodes at euclidean distance less than* r*. Then there is a chain of inclusions*

$$\mathscr{V}_{r_1}(\mathscr{X}) \subset \mathscr{C}_{r_2}(\mathscr{X}) \subset \mathscr{V}_{r_2}(\mathscr{X}) \tag{5.1}$$

whenever

$$\frac{r_2}{r_1} \geq \sqrt{\frac{2d}{d+1}}. \tag{5.2}$$

Thus for ease of computation we shall be using the Vietoris–Rips complex for the simulations contained in this chapter. Theorems 5.7 and 5.10 also assume that the RGSC is constructed using the Vietoris–Rips approach. Although it should be noted that the intuition behind the results also applies to the Čech complex.

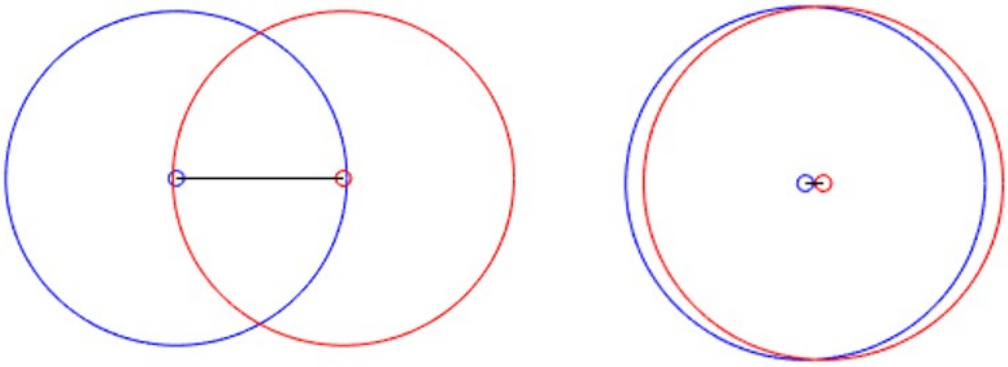
5.3 Creating Holes and Edge Subgraph Centrality

In the previous chapter we saw that edge centralities were not very useful at detecting nodes whose corresponding proteins were essential. However, we now demonstrate that in the 2-dimensional RGSCs they are useful at detecting nodes which we call hole-critical.

Definition 5.6. A node is considered **hole-critical** if its removal would increase the first Betti number, β_1, of the simplicial complex.

That is a node is hole-critical if its removal would introduce a hole.

Remember that in a simplicial complex two edges are adjacent if they have precisely one node in common and they do not take part in the same triangle. Therefore, an edge has high degree if its nodes both have high node degree, and it is not part of many triangles. The interactions and adjacencies in a RGSC are determined by the local geometry which allows us to demonstrate a result about which nodes contribute to the edge degree of a 1-simplex.



(a) (b)

Figure 5.3: The area in which nodes do not contribute to the edge degree is smaller for the longer edge (a) than for the shorter one (b).

Theorem 5.7. *Let S be an RGSC and $\sigma = \{i, j\} \in S$. The degree of σ can be calculated by counting the nodes within the connection radius of precisely one of its two nodes i or j.*

Proof. Recall from Theorem 3.15 that the degree of σ is given by

$$\delta_1(\sigma) = \delta_0(i) + \delta_0(j) - (2 + 2T) \tag{5.3}$$

where δ_k represents the k-degree of a simplex as per Definition 3.14 and T is the number of 2-simplices of which σ is a face.

Let's use the edge from Figure 5.3(a) and let i be the node on the left and j be the node on the right. Given that the circles represent the connection radius of the RGSC then $\delta_0(i)$ would be the total number of nodes which fall within the red circle, $\delta_0(j)$ would be the same for the blue circle and T would be given by the number of nodes which fall within both circles. We can therefore work out the contribution of the 1-simplices which are simplex adjacent to σ through node i to $\delta_1(\sigma)$ by summing all the nodes in the red circle and then subtracting any which are also in the blue circle. We can do the opposite for node j and the two summations are equal to simply summing the number of nodes which are within the connection radius of precisely one of i or j. □

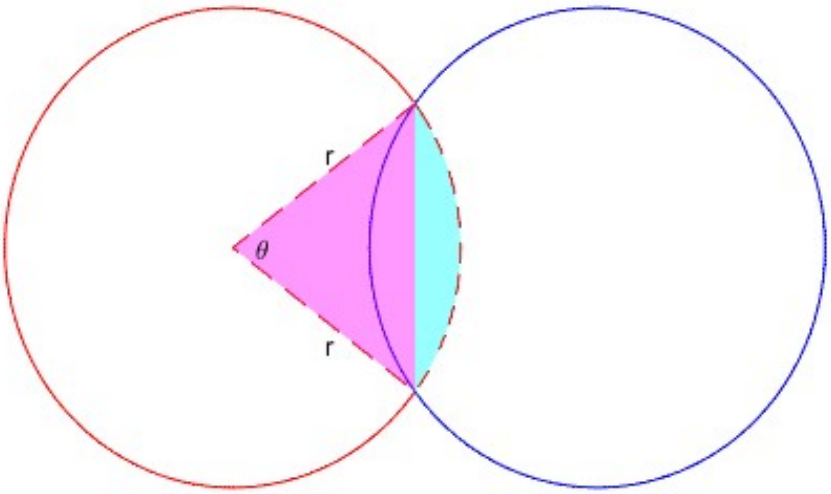


Figure 5.4: A graphical depiction of the areas used to derive the equation of a symmetric lens.

Consider the two edges in Figure 5.3. We would expect that the longer edge would have a higher edge degree because the area where only one of its nodes would be adjacent to a node in that area is larger. That is, there is a larger area contained by the red circle but not the blue circle or by the blue circle but not the red circle for the longer edge than the shorter one.

Before we can prove this assertion we need to calculate the area of intersection of two circles of the same radius, which has a well known formula.

Definition 5.8. A **symmetric lens** is the shape defined by the intersection of two circles of equal radius [91].

Lemma 5.9. *The area of a symmetric lens is* $r^2(\theta - \sin(\theta))$ *where* r *is the radius of the circles and* θ *is the arc length of the intersection of each circle [116, p. 23].*

Theorem 5.10. *Let* S *be an RGSC and* $\sigma = \{i, j\} \in S$ *with* i *and* j *both at least euclidean distance* r*, which is the connection radius of* S*, from the boundary of the area*

on which S is defined. Then the expected degree centrality of σ is given by

$$\mathbb{E}(\delta_1(\sigma)) = 2(n-2)r^2(\pi - (\theta - \sin(\theta))) \tag{5.4}$$

where n is the number of nodes in S and θ is the arc length of the lens created by the overlap of two circles of radius r centred at i and j respectively.

It is also the case that $\mathbb{E}(\delta_1(\sigma))$ increases as the length of σ increases.

Proof. We begin by deriving the formula for $\mathbb{E}(\delta_1(\sigma))$ from the formula for the degree of a 1-simplex given in Theorem 3.15.

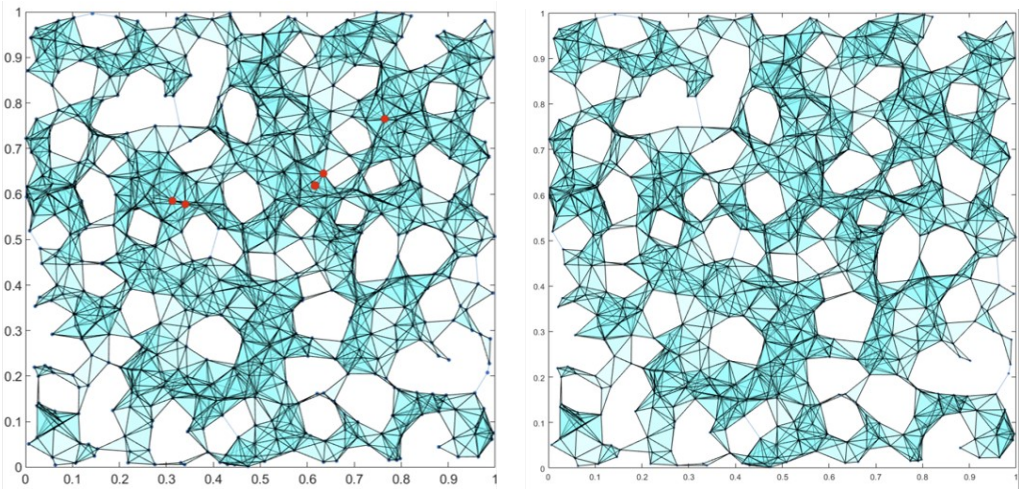
$$\begin{aligned}\mathbb{E}(\delta_1(\sigma_e)) &= \mathbb{E}(\delta_0(i)) + \mathbb{E}(\delta_0(j)) - (2 + 2\mathbb{E}(T)) \\ &= 1 + (n-2)\pi r^2 + 1 + (n-2)\pi r^2 - 2 - 2(n-2)r^2(\theta - \sin(\theta)) \\ &= 2(n-2)r^2(\pi - (\theta - \sin(\theta)))\end{aligned} \tag{5.5}$$

We can assume that the area on which the nodes of the RGSC are scattered is 1. Hence $\mathbb{E}(\delta_0(i)) = 1 + (n-2)\pi r^2$ because we know that node j is adjacent to node i and there are $(n-2)$ other nodes in the RGSC. The proportion of the remaining nodes which we would expect to fall inside the connection radius r is πr^2 because the total area is one.

Similarly, to calculate $\mathbb{E}(T)$ there are $(n-2)$ other nodes and we would expect the proportion of them which would form triangles with the two nodes to be equivalent to the area of the lens, $r^2(\theta - \sin(\theta))$, formed by the intersection of the two of circles of radius, r, centred at i and j divided by the total area the nodes are scattered which we assume, without loss of generality, to be 1.

We can simplify to the expression in the last line of the proof above. As the length of the edge σ_e increases θ decreases and $(\theta - \sin(\theta))$ is a monotonically increasing function of θ. Thus, $(\pi - (\theta - \sin(\theta)))$ is a monotonically decreasing function of θ. Hence, $\mathbb{E}(\delta_1(\sigma_e))$ increases as the length of σ_e increases. □

More intuitively $\delta_1(\sigma)$ is the number of nodes which are within the connection radius of precisely one of the nodes of an edge. Hence the larger this area is the more likely an edge's degree centrality is high. This area is larger when the edge is longer as can be seen from Figure 5.3.



(a) (b)

Figure 5.5: The nodes highlighted are the most edge subgraph central ones for this RGSC (a) we can see the effect of removing these nodes from the RGSC(b).

The subgraph centrality of a node in a network has been shown to be a measure of the number of subgraphs of a graph that a node participates in. That is a node has a high subgraph centrality if it not only has a high degree but many of the nodes that it is connected to also have high degree and have many connections between each other. Therefore, in a simplicial complex an edge is likely to have high subgraph centrality if it is not only of high degree but if it is also adjacent to many edges of high degree. On random geometric simplicial complexes we have demonstrated that longer edges have higher degree so an edge is likely to have higher subgraph centrality if it is long and it has at least one node which is a member of many other long edges. Additionally, the other end points of these edges would need to be far away from each other to minimise the chances of forming triangles. If this situation arose then all of the other edges surrounding the same node would also have high subgraph centrality. So a node which is involved in many edges which have high subgraph centrality is likely to be at one end of many long edges and therefore quite far from other nodes in the euclidean space the nodes were spread on. That is such a node is likely to be hole–critical.

We can connect this idea to the case of wireless sensor networks and the problem of sensors failing with the result of a gap appearing in the coverage of the WSN. Therefore,

Betti Number Change	1 node	2 nodes	3 nodes	4 nodes	5 nodes
1	122	209	222	189	186
2	3	36	79	138	155
3	0	3	9	24	45
4	0	0	0	3	13
5	0	0	0	0	1
-1	0	0	0	3	2
Total	125	248	310	357	402
Unchanged	375	252	190	143	98

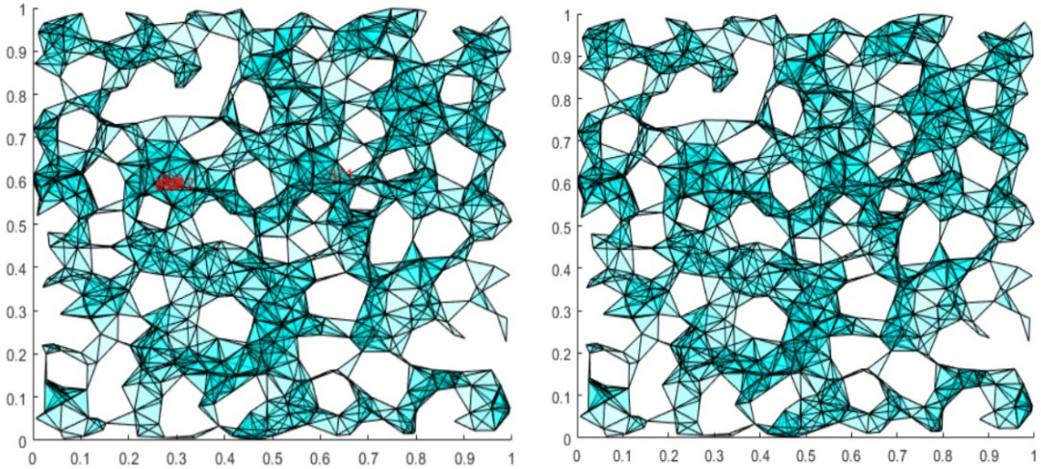
Table 5.1: 500 RGSCs were generated with connection radius 0.07 and 750 nodes. The number of these RGSCs which have changed Betti number by a given value after deleting a certain number of nodes. Node to be deleted is picked according to the subgraph centralities of their edges.

we now consider the 100 most subgraph central edges in an RGSC and whichever node is contained most often in this list of edges will be considered the most edge subgraph central node.

We demonstrate experimentally the effectiveness of removing the most edge subgraph central node of a network as a way of introducing holes to a network. Figure 5.2 is an RGSC with sidelength 1, connection radius 0.07 and $n = 750$ nodes. It has a first Betti number $\beta_1 = 45$ which is equivalent to saying that the number of gaps in coverage of a WSN being modelled by this RGSC is 45. In Figure 5.5 (a) we have highlighted the five most edge subgraph central nodes of this RGSC as determined by the following process:

1. Calculate edge subgraph centrality for every edge in the simplicial complex.
2. Identify the 100 edges with the largest edge subgraph centralities.
3. Calculate which node features in the largest number of these highly edge subgraph central edges. If there is a tie then break the tie by picking whichever of the tied nodes is contained in an edge closest to the top of the list.
4. Remove this node
5. Repeat process on reduced simplicial complex until 5 nodes have been removed.

When these nodes are removed then the resulting simplicial complex is that shown in Figure 5.5 (b). Note that it has a Betti number of 46 due to the hole which has opened in the region close to $(0.3, 0.6)$. Moreover, the hole at $(0.7, 0.65)$ has increased



(a) (b)

Figure 5.6: The nodes highlighted are the most node subgraph central ones for this RGSC (a) we can see the effect of removing these nodes from the RGSC(b).

in size further disrupting communication between the nodes in this area.

This process was repeated on 500 different RGSCs with connection radius 0.07 and $n = 750$ nodes. The average number of edges of these RGSCs was 4071 with a standard deviation of 74.4, a minimum of 3856 and a maximum of 4310 which means that roughly the top 2.5% of edges contribute to the selection of which node to delete. Table 5.1 details the changes to the Betti numbers of these RGSCs throughout this process. After deletion of just five nodes 80.4% of the RGSCs had seen β_1 increase by at least one which corresponds to finding at least one hole critical node and in 42.8% of cases at least two hole-critical nodes had been found.

To further demonstrate that this result is not just a normal result of deleting nodes three further node deletion strategies have been tried. These strategies are deleting edges according to node closeness centrality, node subgraph centrality and random deletion. We discuss each strategy in turn.

As previously mentioned, nodes have high subgraph centrality when they have high degree and they are adjacent to many other nodes which also have this property. Therefore, on an RGSC we would expect that nodes with high subgraph centrality would be found in dense clusters. These nodes would be unlikely to be hole-critical because there are likely to be many other nodes nearby which would also need to be deleted to induce

Betti Number Change	1 node	2 nodes	3 nodes	4 nodes	5 nodes
1	7	21	28	49	60
2	0	0	0	1	2
Total	7	21	28	50	62
Unchanged	493	479	472	450	438

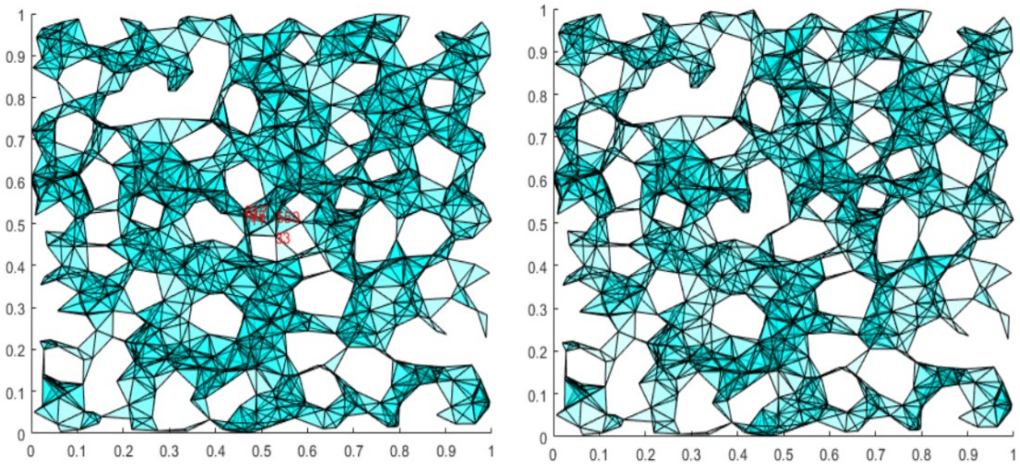
Table 5.2: 500 RGSCs were generated with connection radius 0.07 and 750 nodes. The number of these RGSCs which have changed Betti number by a given value after deleting a certain number of nodes. Node to be deleted is picked according to their subgraph centrality.

a hole in that location which was confirmed by experimentation. The same process was used as the experiment where nodes were deleted according to edge subgraph centrality but instead the node with the largest node subgraph centrality was iteratively removed. Table 5.2 details that deletion of 5 nodes resulted in an increased Betti number in only 12.4% of the RGSCs. We can also see this process play out in Figure 5.6 where the most node subgraph central nodes are highlighted and then removed from the RGSC from Figure 5.2. This process has no effect on the number of holes in the RGSC.

On a RGSC the nodes with the highest node closeness centralities are found in the centre of the space on which the nodes have been uniformly distributed. When the most central nodes are deleted then we have a situation where one large hole forms in the centre of the RGSC. Table 5.3 shows that a large number of simplicial complexes change Betti number (many more than the case of node centrality but less than in the case of edge subgraph centrality) and a larger portion of the changes in Betti number are by one even after deletion of multiple nodes. However, there are more decreases of Betti number with closeness than there are for edge subgraph centrality. These decreases are due to multiple holes in the middle of the RGSC merging together to form one large hole. This process plays out in the case of the RGSC from Figure 5.2, which is depicted in Figure 5.7. The five small holes in the centre of the square join to form three larger ones after removing the nodes with the largest closeness.

Finally, to provide a control to these experiments a series of five nodes were deleted randomly from 500 RGSCs and the resulting changes in the Betti Number were tracked. The results are detailed in 5.4.

We have four different node deletion strategies which have four different effects on



(a) (b)

Figure 5.7: The nodes highlighted are the most node closeness central ones for this RGSC (a) we can see the effect of removing these nodes from the RGSC(b).

Betti Number Change	1 node	2 nodes	3 nodes	4 nodes	5 nodes
1	92	118	141	149	160
2	1	13	23	34	47
3	0	1	2	3	5
4	0	0	2	3	3
-1	25	45	61	72	86
-2	8	13	22	28	26
-3	0	0	2	4	11
-4	0	0	0	1	2
Total	126	190	253	294	340
Unchanged	374	310	247	206	160

Table 5.3: 500 RGSCs were generated with connection radius 0.07 and 750 nodes. The number of these RGSCs which have changed Betti number by a given value after deleting a certain number of nodes. Node to be deleted is picked according to highest closeness centrality.

Betti Number Change	1 node	2 nodes	3 nodes	4 nodes	5 nodes
1	34	51	67	87	104
2	1	2	9	16	17
3	0	0	0	1	0
4	0	0	0	0	1
-1	28	53	68	68	74
-2	1	4	7	11	14
-3	0	0	0	1	1
Total	64	110	151	184	211
Unchanged	436	390	349	316	289

Table 5.4: 500 RGSCs were generated with connection radius 0.07 and 750 nodes. The number of these RGSCs which have changed Betti number by a given value after deleting a certain number of nodes. Node to be deleted is picked randomly.

the number of holes in the RGSC. We have the random behaviour exhibited by the random deletion, the behaviour of creating lots of small holes exhibited by the edge subgraph centrality, the behaviour of creating a large central hole exhibited by the closeness centrality and the behaviour of having little effect on the Betti number of the node subgraph centrality. The results of these experiments are summarised in 5.5 for ease of the reader.

We have demonstrated that the edge subgraph centrality can be used to detect hole-critical nodes in RGSCs which has implications for the study of WSNs because it means that when nodes fail in a WSN the implications for which other nodes have become hole-critical can be calculated quickly to allow movement of resources towards regions where gaps in coverage are more likely to appear. This algorithm could be combined with other algorithms in the management of WSNs.

Betti Number Change	1 node	2 nodes	3 nodes	4 nodes	5 nodes
Edge Sub +	125	248	310	354	400
Edge Sub -	0	0	0	3	2
Node Sub +	7	21	28	50	62
Node Sub -	0	0	0	0	0
Closeness +	93	132	168	189	215
Closeness -	33	58	85	105	125
Random +	35	53	76	104	122
Random -	29	57	75	80	89

Table 5.5: This table summarises the previous experiments. The four methods of node deletion Edge Sub (Largest number of nodes with high edge subgraph centrality), Node Sub (Largest Node Subgraph Centrality), Closeness (Largest closeness centrality), Random are accompanied by a sign + or -. The + indicates a positive change in Betti Number, the - indicates a negative change. The numbers are the number of RGSCs which displayed a change of this type after deletion of the given number of nodes.

Chapter 6

Epidemic Spreading on Random Rectangular Annular Graphs

6.1 Literature Review

When Gilbert introduced his random planar graphs [63] he also mentioned the spread of infectious diseases which is only really relevant to certain settings, in plants for example [7, 47]. In humans or other mobile animals this strategy is less likely to be appropriate because the points are unlikely to remain stationary long enough for transmission due to their location to become the main cause of transmission. However, when we consider a system of plants in a field, each plant could pass on a disease to other plants close to it but cannot pass on the disease to those plants that are further away. The appropriateness of modelling the spread of plant disease with random geometric graphs is, of course, dependent on the mode of disease transmission. Many aspects of disease transmission on random geometric graphs have been studied including the work on how far a disease can spread when the connection radius of the RGG is varied by Saha et al. who calculated both the geographical limit of how far the disease could spread and the expected number of individuals infected under an SIR model [136]. Another example is that Panicker and Sasidevan were able to produce the oscillations in prevalence that characterised the COVID-19 epidemic using evolving random geometric graphs where the density of nodes in the graph was varied depending on the prevalence

of disease which mimicked the introduction and relaxation of social distancing measures during the early stages of the outbreak [117].

In recent years there has also been an explosion of interest in extending the study of random geometric graphs to consider shapes other than squares, partly due to an appreciation that not all fields are square and the shape of a field may change the properties under study. One example is the work by Martínez-Martínez et al. who looked at the random circular graph with the nodes distributed from the centre based on the normal distribution [101]. They looked at how properties of the graph changed under different values for the standard deviation of the normal distribution [101]. Allen-Perkins studied the average degree, degree distribution and clustering coefficient among other properties of random geometric graphs on the surface of a sphere [2]. The random rectangular graph was introduced by Estrada and Sheerin in 2015 [51] while Estrada et al. studied epidemic spreading over random rectangular graphs [47], Estrada and Chen looked at Synchronizability [45] and Estrada and Sheerin investigated consensus dynamics [52]. Later Alonso et al. added a random weighting component to both the nodes and edges of the random rectangular graph and then studied the spectral properties of their construction [3] and Arias et al. expanded on epidemic spreading on random rectangular graphs adding in long range dispersal through Mellin and Laplace transforms which adds the wind based dispersal of fungal spores to the model to complement the random rectangular graph's ability to model short range dispersal of these spores by insects or mammals [7]. Deshpande used random rectangular graphs as part of his study of range expansion of species noting that range expansion models typically feature long and narrow areas [35]. These studies point to the fact that the location of the boundaries of a random rectangular graph affect the dynamics of processes on it and clearly not all locations which interest researchers are square.

Additionally, not all locations of interest are whole, for example if one was attempting to model the spread of plant disease across Central Park in Manhatten the most appropriate shape would be a rectangle with a large hole in the middle to factor in the presence of the lake on which many plant species would be unable to survive. If we were modelling the spread of disease through populations of plants in the Scottish

Highlands, or a similarly rugged terrain, we would wish to understand what effect the lack of plants on the tops of mountains or in the middle of lochs would have on the dynamics. Giles et al. considered annuli in 2016 [64] establishing the connectivity of Random Geometric Graphs over Annuli. Angel and Spinka considered random geometric graphs on circles in 2021 [6] and Galhotra et al. investigated connectivity of random geometric graphs on spheres of different dimensions of which the one dimensional case is an annulus [58]. The previous chapter explored various methods of deleting nodes from an RGSC and the effect that has on their homology. This chapter investigates how the homology of the underlying space affects the epidemic spreading on random geometric graphs scattered on them which is of relevance when analysing the spread of a plant disease in a wild or domestic setting with large areas where certain plants would not grow such as mountains, lakes or uncultivated areas surrounding telegraph poles. Note that large in this instance is relative to the expected maximum distance between plants which could result in the transmission of a disease.

6.2 Expected Average Degree

Estrada and Sheerin [51, 144] showed that the expected node degree of a node, v_i, in a random rectangular graph is given by $\mathbb{E}(k_i) = \frac{(n-1)A_i}{ab}$ where a and b are the side lengths of the rectangular area and A_i is the area both within a radius r of v_i and also inside the rectangle. They averaged over every position in the rectangle to get $\mathbb{E}(\overline{k}) = \frac{(n-1)\int_p A_p}{(ab)^2}$ where $\int_p A_p$ is the summation of A_i for all i i.e. for every possible point in the rectangle. We now generalise their result to rectangular areas with holes.

Definition 6.1. Let there be a rectangular area with side lengths a, b with another rectangular area in the middle with side lengths R_1, R_2. Randomly and uniformly scatter n nodes in the area inside the larger rectangle but outside the smaller one. Then as usual connect two nodes if they are within distance r of each other. This construction is a **random rectangular annular graph**.

Note that it is not required that the edges of the inner rectangle are parallel to the edges of the outer rectangle. All theorems in this chapter are accurate regardless of the

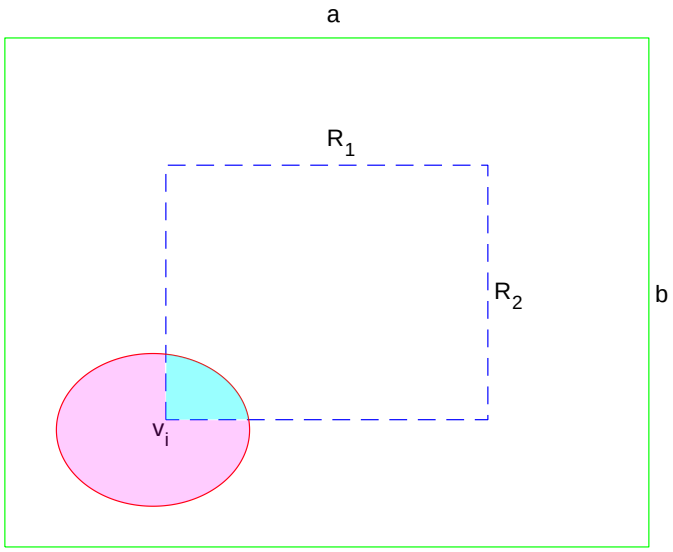


Figure 6.1: A graphical look at how the area within the inner rectangle but outwith the outer rectangle can be calculated.

orientation of the inner rectangle.

As in the case of the random rectangular graphs the expected node degree of a node v_i is given by the number of other nodes $(n-1)$ multiplied by the fraction of the area within which nodes can be scattered which is within distance r from v_i. The difference this time is that the circular area within which all nodes share an edge with v_i may be interrupted by the internal boundary as well as the external one (see Figure 6.1).

We can let the area on which nodes can be scattered within radius r of a node v_i be denoted by B_i. The total area on which nodes can be scattered is given by $(ab - R_1R_2)$. Then we can follow the example of before and say that the expected node degree of a node v_i is given by

$$\mathbb{E}(k_i) = \frac{(n-1)B_i}{ab - R_1R_2}. \tag{6.1}$$

From Figure 6.1 we can see that the red area, B_i, can also be expressed as the whole coloured area (which shall be called A_i) with the blue area subtracted. The blue area is the area within radius r of a node v_i which is within the smaller rectangle and will be denoted C_i. Then we can express B_i as $A_i - C_i$ and therefore have that

$$\mathbb{E}(k_i) = \frac{(n-1)(A_i - C_i)}{ab - R_1 R_2}. \tag{6.2}$$

We can average over all points in the area on which nodes are spread to get

$$\mathbb{E}\left(\overline{k}\right) = \frac{(n-1)\left(\int_p A_p - C_p\right)}{(ab - R_1 R_2)^2} = \frac{(n-1)\left(\int_p A_p - \int_p C_p\right)}{(ab - R_1 R_2)^2}. \tag{6.3}$$

Note that $\int_p A_p$ and $\int_p C_p$ are the sum, across all points in the area of the RRAG, of the area which is within r of a given point p and also within the outer rectangle (the coloured area in Figure 6.1) and the sum of the area which is within r of a given point p and also within the inner rectangle (the blue area in Figure 6.1) respectively. The summations of these contributions are then averaged across the whole area of the RRAG by division by $(ab - R_1 R_2)$ which we can see is a squared term in Equation (6.3) as opposed to the singular instance of it in Equation (6.2).

Theorem 6.2. *For a Random Rectangular Annular Graph where we assume that the inner area is a distance at least r from the edge of the larger area and that $R_1, R_2 > r$, the expectation of the average node degree of the graph is given by:*

$$\mathbb{E}\left(\overline{k}\right) = \frac{(n-1)\left(\pi r^2 (ab - R_1 R_2) - \frac{4}{3}(a+b) r^3 + \frac{1}{2} r^4 - 4r^3 \left(\frac{R_1}{3} + \frac{R_2}{3} - \frac{r}{8}\right)\right)}{(ab - R_1 R_2)^2}. \tag{6.4}$$

Proof. 1. Overview

We consider quarter circles around each point and multiply the results by 4 at the end. Note that the theorem is stated such that it is not necessary for the sides of the inner rectangle to be parallel to the sides of the outer rectangle because we consider the interaction with the inner rectangle separately from the other contributions. Therefore we could reorient the RRAG between the calculation of $\int_p A_p$ and $\int_p C_p$ and retrieve the expected result. Additionally, any point on the boundary of the inner rectangle being at least r from the boundary of the outer rectangle means that a quarter circle which interacts with the inner boundary cannot also overlap with the outer boundary.

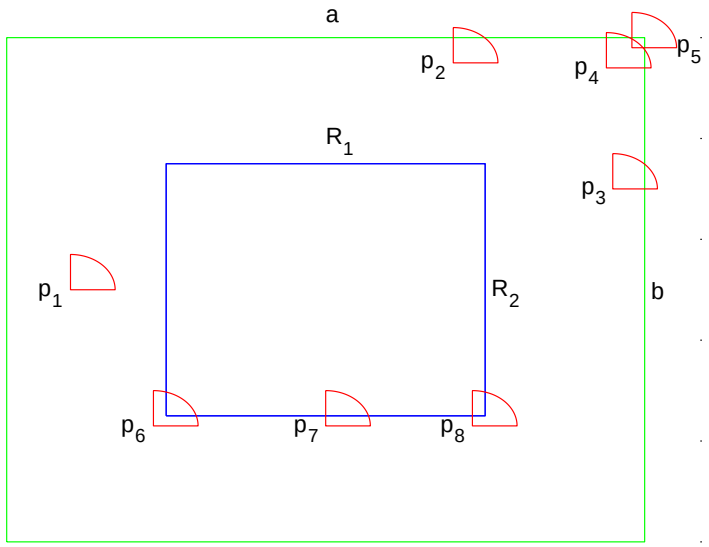


Figure 6.2: The different possible cases in terms of the nodes interacting with the boundary of either the inner or the outer rectangle.

We can refer to Figure 6.2 which shows us that there are 8 different ways in which the area within r of a point could interact with either the inner or outer rectangles. The simplest case is that of p_1 which does not interact with the edges at all. We then have that p_2 and p_3 interact with precisely one of the boundaries of the outer rectangle while p_4 and p_5 interact with two boundaries of the outer rectangle. We also have that p_6, p_7 and p_8 interact only with the inner boundary.

This proof is divided into parts, of which this overview is the first, to aid readability. The calculation $\mathbb{E}(\overline{k}) = \frac{(n-1)\left(\int_p A_p - \int_p C_p\right)}{(ab - R_1 R_2)^2}$ has been separated into $\int_p C_p$ which deals with the interaction with the inner boundary and $\int_p A_p$ which will deal with everything else. The second part will deal with the calculation of $\int_p A_p$ for points like p_1, p_6, p_7 and p_8 which do not intersect the outer rectangle and so can be treated identically for this evaluation. The third part will consider the calculation of $\int_p A_p$ for points like p_2 and p_3 while the fourth will evaluate the contribution of points like p_4 and p_5 to $\int_p A_p$. The fifth part will put the results of parts 2, 3 and 4 together to get a final evaluation for $\int_p A_p$. The sixth part moves on to calculate the contribution of points similar to p_6 to $\int_p C_p$ remembering that for point similar to p_1, p_2, p_3, p_4 and p_5 their contribution to $\int_p C_p$ is 0.

The seventh part gives us a useful shortcut for the calculation of the contribution of points similar to p_8 to $\int_p C_p$ while the eighth and ninth parts determine the contribution of points similar to p_7 and p_8 to $\int_p C_p$. The final part brings the contributions of the different kinds of points to $\int_p C_p$ together and then delivers the final formula for $\mathbb{E}\left(\overline{k}\right)$. Some of these parts are further divided into sub parts in order to consider different terms of each integration.

2. Calculation of $\int_p A_p$ for points similar to p_1, p_6, p_7 or p_8.

   The area of the quarter circle at p_1, p_6, p_7 or p_8 can be calculated as $\frac{\pi r^2}{4}$ and any point which is at least r from both the top and right edges of the outer rectangle has this form. Therefore we can calculate that the contribution to $\int_p A_p$ from points of the type p_1, p_6, p_7 or p_8 is

$$\frac{\pi r^2((a-r)(b-r) - R_1R_2)}{4}. \tag{6.5}$$

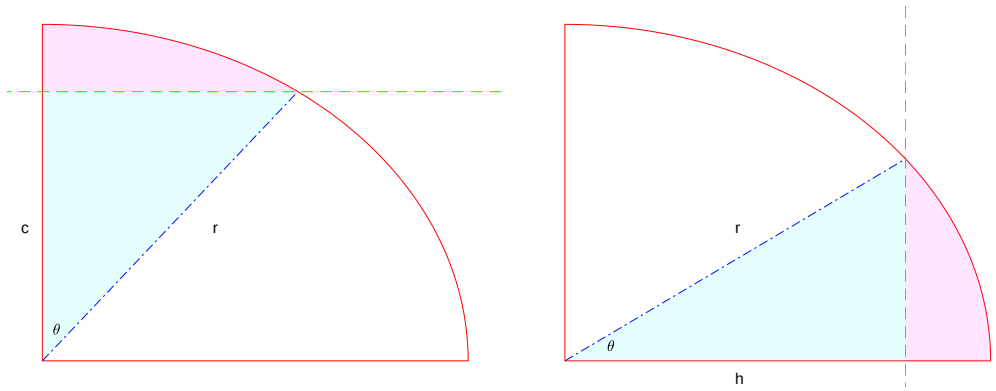
3. Calculation of $\int_p A_p$ for points similar to p_2 or p_3.

   We can now turn our attention to points of the form p_2 and p_3. A close up of these two situations is displayed in Figure 6.3. To work out the area contained within the rectangle for p_2 we need to subtract the red area above the green dashed line from the area of the quarter circle, $\frac{\pi r^2}{4}$. We can work out the red area as the area of the coloured circular sector with the area of the blue triangle below the green dashed line subtracted.

   The area of a circular sector is given by $\frac{r^2\theta}{2}$ with $\theta = \cos^{-1}\left(\frac{c}{r}\right)$ and the area of the blue triangle is given by $\frac{c\sqrt{r^2-c^2}}{2}$. We can put all of these formulae together to get the following formula for the area inside the rectangle for a point like p_2

$$\frac{\pi r^2}{4} - \frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right) + \frac{c\sqrt{r^2-c^2}}{2}. \tag{6.6}$$

   We can use a similar process to define a result for a point similar to p_3 as



(a) (b)

Figure 6.3: A node within r of the top of the outer rectangle (a) and a node within r of the right edge of the outer rectangle (b). In both cases the edge of the rectangle is depicted by the green dashed line.

$$\frac{\pi r^2}{4} - \frac{r^2}{2}\cos^{-1}\left(\frac{h}{r}\right) + \frac{h\sqrt{r^2-h^2}}{2}. \tag{6.7}$$

We can integrate this equation from 0 to r to get the sum of all the points on a line descending perpendicularly from the top or right hand side of the rectangle of length r. We can then multiply this result by $(a-r)+(b-r)$ to get the contribution from points similar to p_2 and p_3. To calculate

$$\int_0^r \frac{\pi r^2}{4} - \frac{r^2}{2}\cos^{-1}\left(\frac{h}{r}\right) + \frac{h\sqrt{r^2-h^2}}{2}dh \tag{6.8}$$

we work out each term separately and then put them together again.

(a) Integral of $\int_0^r \frac{\pi r^2}{4}dh$ term in Equation (6.8).

We start with

$$\int_0^r \frac{\pi r^2}{4}dh = \left[\frac{\pi r^2 h}{4}\right]_0^r = \frac{\pi r^3}{4}. \tag{6.9}$$

(b) Integral of $\int_0^r \frac{r^2}{2}\cos^{-1}\left(\frac{h}{r}\right)dh$ term in Equation (6.8).

We have

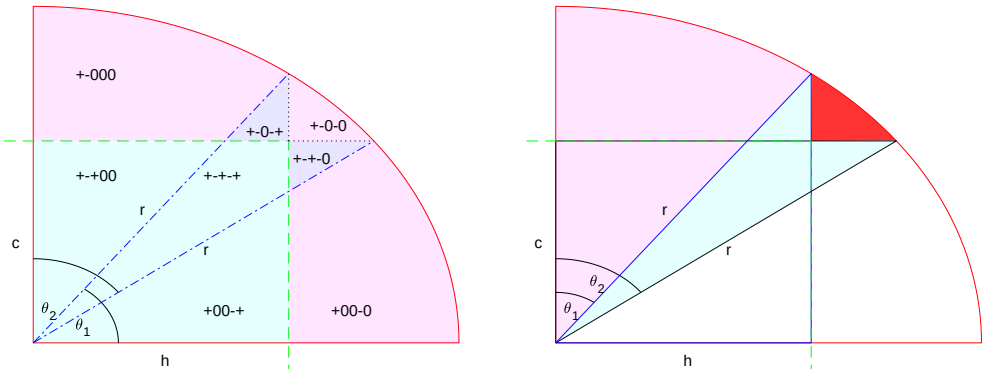


Figure 6.4: The different regions and the effect of addition and subtraction of each term in Equation 6.13 (a) and a depiction of how to calculate the dark red area as explained by Equation 6.15 (b).

$$\int_0^r \frac{r^2}{2}\cos^{-1}\left(\frac{h}{r}\right)dh = \left[\frac{r^2h}{2}\cos^{-1}\left(\frac{h}{r}\right) - \frac{r^3}{2}\sqrt{1-\left(\frac{h}{r}\right)^2}\right]_0^r = \frac{r^3}{2} \tag{6.10}$$

by substitution of $g = \frac{h}{r}$ followed by evaluation of $\int \cos^{-1}(g)dg$ by integration by parts.

(c) Integral of $\int_0^r \frac{h\sqrt{r^2-h^2}}{2}dh$ term in Equation (6.8).

For our final term

$$\int_0^r \frac{h\sqrt{r^2-h^2}}{2}dh = \left[-\frac{(r^2-h^2)^{\frac{3}{2}}}{6}\right]_0^r = \frac{r^3}{6} \tag{6.11}$$

by substitution of $f = r^2 - h^2$.

(d) Result of $\int_p A_p$ for points similar to p_2 or p_3.

We can now put together our results from Equations (6.9), (6.10) and (6.11) to get the contribution to $\int_p A_p$ by points similar to p_2 and p_3 to be

$$(a+b-2r)\left(\frac{\pi r^3}{4} - \frac{r^3}{3}\right). \tag{6.12}$$

4. Calculation of $\int_p A_p$ for points similar to p_4 or p_5

(a) Description of approach.

We can now move on to consider points which are within r of a corner in terms of both the x-axis and the y-axis i.e. points which are similar to p_4 or p_5. Notice that in the case of p_4 we can just add together the contributions from the parts which are similar to p_2 and p_3 and integrate over c and h respectively. We can calculate this area as

$$\frac{\pi r^2}{4} - \frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right) + \frac{c\sqrt{r^2-c^2}}{2} - \frac{r^2}{2}\cos^{-1}\left(\frac{h}{r}\right) + \frac{h\sqrt{r^2-h^2}}{2}. \tag{6.13}$$

However, this approach gives us a problem when we try to use it to analyse points similar to p_5 i.e. a point which is within a euclidean distance less than r from the corner. That is for any point where $r < \sqrt{c^2+h^2}$.

It is useful to have a graphical understanding of how each separate area is affected by each term of Equation (6.13) which is provided in Figure 6.4 (a) where each region has a series of five symbols from $\{+,-,0\}$. A + symbol as the first entry of the series indicates that this region is part of the area that is added as part of the first term of Equation (6.13) and a − symbol as the second entry indicates that this region is subtracted as part of the second term of Equation (6.13) i.e. the $-\frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right)$ term. The effect of the third, fourth and fifth entries Equation (6.13) are similarly indicated on Figure 6.4 (a).

We wish to calculate the area of the blue rectangle and so we would like that any point inside the light blue rectangle has one more addition than subtraction in Equation (6.13) which is the case. However, we would also like that any point outside of the blue rectangle has an equal number of additions and subtractions and this requirement is not met because the region which is depicted in dark red in Figure 6.4 (b) has been subtracted from Equation (6.13) twice but only added once.

Our approach is to integrate Equation (6.13) from 0 to r with respect to h and then integrate the result from 0 to r with respect to c. Then to work

out a formula for the area of the dark red region from Figure 6.4 (b) for a given c, h with $r < \sqrt{c^2 + h^2}$. Integration of this formula from $\sqrt{r^2 - c^2}$ to r with respect to h and then integrating the result from 0 to r with respect to c gives the value which has to be added back to the integration of Equation (6.13) to compensate for the extra subtraction of the dark red region which happens for points similar to p_5.

(b) Integration of Equation (6.13) in calculation of $\int_p A_p$ for points similar to p_4 or p_5.

We start with the integration of Equation (6.13) from r to 0 with respect to h and from r to 0 with respect to c. We can use the results of Equations (6.9), (6.10) and (6.11) to get

$$\begin{aligned}
&\int_0^r \int_0^r \frac{\pi r^2}{4} - \frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right) + \frac{c\sqrt{r^2-c^2}}{2} - \frac{r^2}{2}\cos^{-1}\left(\frac{h}{r}\right) + \frac{h\sqrt{r^2-h^2}}{2} dc dh \\
=&\int_0^r \frac{\pi r^3}{4} - \frac{r^3}{2} + \frac{r^3}{6} - \frac{r^3}{2}\cos^{-1}\left(\frac{h}{r}\right) + \frac{rh\sqrt{r^2-h^2}}{2} dh \\
=&\frac{\pi r^4}{4} - \frac{r^4}{2} + \frac{r^4}{6} - \frac{r^4}{2} + \frac{r^4}{6} \\
=&r^4\left(\frac{\pi}{4} - \frac{4}{6}\right).
\end{aligned} \tag{6.14}$$

(c) Derivation of formula for the area of the dark red region in Figure 6.4 (b).

In Figure 6.4 (b) the area of the whole coloured area can be calculated as $\frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right)$ as before because $\theta_2 = \cos^{-1}\left(\frac{c}{r}\right)$. The red circular sector can be removed to leave just the blue circular sector. The area of this red sector is given by $\frac{r^2}{2}\sin^{-1}\left(\frac{h}{r}\right)$ as $\theta_1 = \sin^{-1}\left(\frac{h}{r}\right)$.

We can then subtract the triangle with the blue outline and the one with the black outline which have areas $\frac{h\sqrt{r^2-h^2}}{2}$ and $\frac{c\sqrt{r^2-c^2}}{2}$ respectively. The result is that the only area of the blue sector which remains is that which is coloured dark red. However this subtraction also results in the area that is within the rectangle being subtracted as well. Once we add this area in we are left with the following equation for the dark red area

$$\frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right)-\frac{r^2}{2}\sin^{-1}\left(\frac{h}{r}\right)-\frac{h\sqrt{r^2-h^2}}{2}-\frac{c\sqrt{r^2-c^2}}{2}+ch. \tag{6.15}$$

(d) Integration of Equation (6.15) in calculation of $\int_p A_p$ for points similar to p_4 or p_5.

The following equation describes the value that must be added back into Equation (6.14) to account for the subtraction of the dark red region from Figure 6.4 (b) in that result

$$\int_0^r\int_0^{\sqrt{r^2-c^2}}\frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right)-\frac{r^2}{2}\sin^{-1}\left(\frac{h}{r}\right)-\frac{h\sqrt{r^2-h^2}}{2}-\frac{c\sqrt{r^2-c^2}}{2}+ch\ dcdh. \tag{6.16}$$

As with integrals for the other points we will do this calculation term by term. We begin with the parts which do not depend on h or are linear in terms of it.

i. Integration of $\frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right)-\frac{c\sqrt{r^2-c^2}}{2}+ch$ term in calculation of Equation (6.15).

The equation below describes the results of the the first stage of the integration of the first, fourth and fifth terms of Equation (6.15)

$$\begin{aligned}
&\int_0^r\int_0^{\sqrt{r^2-c^2}}\frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right)-\frac{c\sqrt{r^2-c^2}}{2}+ch\ dhdc\\
=&\int_0^r\left[\frac{hr^2}{2}\cos^{-1}\left(\frac{c}{r}\right)-\frac{hc\sqrt{r^2-c^2}}{2}+\frac{ch^2}{2}\right]_0^{\sqrt{r^2-c^2}}dc\\
=&\int_0^r\frac{r^2\sqrt{r^2-c^2}}{2}\cos^{-1}\left(\frac{c}{r}\right)dc\\
=&\frac{r^4}{8}(\pi^2-1).
\end{aligned} \tag{6.17}$$

The step from the third line to the fourth line is done by using two substitutions, $u=\cos^{-1}\left(\frac{c}{r}\right)$ and $v=\frac{c}{r}$. We will put this result together with the results of the other terms later.

ii. Integration of $\frac{h\sqrt{r^2-h^2}}{2}$ term in calculation of Equation (6.15).

We can use Equation (6.11) to get

$$\begin{aligned}
&\int_0^r \int_0^{\sqrt{r^2-c^2}} -\frac{h\sqrt{r^2-h^2}}{2} dhdc \\
&= -\int_0^r \left[-\frac{(r^2-h^2)^{\frac{3}{2}}}{6} \right]_0^{\sqrt{r^2-c^2}} dc \\
&= -\int_0^r \frac{r^3}{6} - \frac{c^3}{6} = -\left(\frac{r^4}{6} - \frac{r^4}{24} \right) = -\frac{r^4}{8}.
\end{aligned} \tag{6.18}$$

iii. Integration of $\sin^{-1}\left(\frac{h}{r}\right)$ term in calculation of Equation (6.15).

Finally we can analyse the second term from Equation (6.15). We can evaluate

$$\begin{aligned}
&\int_0^r \int_0^{\sqrt{r^2-c^2}} -\frac{r^2}{2} \sin^{-1}\left(\frac{h}{r}\right) dhdc \\
&= -\int_0^r \frac{r^2\sqrt{r^2-c^2}}{2} \sin^{-1}\left(\frac{\sqrt{r^2-c^2}}{r}\right) + \frac{r^2 c}{2} - \frac{r^3}{2} dc \\
&= -\frac{r^4}{8}(\pi^2 - 1) + \frac{r^4}{4}.
\end{aligned} \tag{6.19}$$

The first step can be completed by substitution of $g = \frac{h}{r}$ followed by integration by parts using $u = \sin^{-1}(g)$ and $dv = dg$. The final step is a substitution of $f = 1 - g^2$.

The first term in the second step can be evaluated to be $-\frac{r^4}{8}(\pi^2 - 1)$ through the following series of substitutions

$$\begin{aligned}
u &= \sin^{-1}\left(\frac{\sqrt{r^2-c^2}}{r}\right) \\
v &= \frac{\sqrt{r^2-c^2}}{r} \\
s &= r^2 - c^2.
\end{aligned} \tag{6.20}$$

The second and third terms can be evaluated to be $\frac{r^4}{4}$.

iv. Final evaluation of Equation (6.15).

We are now in a position to simplify the expression in Equation (6.15) by combining the results from Equations (6.17), (6.18) and (6.19) together to get

$$\begin{aligned}&\int_0^r \int_0^{\sqrt{r^2-c^2}} \frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right) - \frac{r^2}{2}\sin^{-1}\left(\frac{h}{r}\right) + \\ &\int_0^r \int_0^{\sqrt{r^2-c^2}} -\frac{h\sqrt{r^2-h^2}}{2} - \frac{c\sqrt{r^2-c^2}}{2} + ch \; dhdc \\ =&\frac{r^4}{8}(\pi^2-1) - \frac{r^4}{8} - \frac{r^4}{8}(\pi^2-1) + \frac{r^4}{4} = \frac{r^4}{8}.\end{aligned} \tag{6.21}$$

(e) Final Result of $\int_p A_p$ for points similar to p_4 or p_5.

Recall that we had calculated the contribution to $\int A_p$ by points similar to p_4 or p_5 in Equation (6.14) but that for points similar to p_5 this calculation required the dark red region from Figure 6.4 (b) to be added back in. The value of this region for all points such that $r < \sqrt{c^2+h^2}$ was calculated in Equation (6.21) which means we can now deduce the contribution to $\int A_p$ by points that are similar to p_4 or p_5. This equation gives the value

$$r^4\left(\frac{\pi}{4} - \frac{4}{6}\right) + \frac{r^4}{8} = r^4\left(\frac{\pi}{4} - \frac{13}{24}\right). \tag{6.22}$$

5. Final Result of $\int A_p$.

We can now work out $\int A_p$ by combining the results from Equation (6.5), which calculated the contribution from points similar to p_1, p_6, p_7 and p_8, Equation (6.12), which calculated the contribution from points similar to p_2 and p_3, and Equation (6.22), which calculated the contribution from points similar to p_4 and p_5. We can then multiply by 4 to get the final value for $\int A_p$,

$$
\begin{aligned}
&\int A_p \\
=&\pi r^2((a-r)(b-r) - R_1R_2) + 4(a+b-2r)r^3\left(\frac{\pi}{4} - \frac{1}{3}\right) + 4r^4\left(\frac{\pi}{4} - \frac{13}{24}\right) \\
=&\pi r^2(ab - R_1R_2) - (a+b)\pi r^3 + \pi r^4 + (a+b)\pi r^3 - \frac{4}{3}(a+b)r^3 - 2\pi r^4 \\
&+ \frac{8}{3}r^4 + \pi r^4 - \frac{52}{24}r^4 \\
=&\pi r^2(ab - R_1R_2) - \frac{4}{3}(a+b)r^3 + \frac{1}{2}r^4.
\end{aligned} \tag{6.23}
$$

Note that if we set $R_1, R_2 = 0$ we would recover the formula discovered by Estrada and Sheerin [51, 144] for the case that $r \leq b$. The restriction that $R_1, R_2 > r$ means that we are never in any of the other regions defined by them, $\left(b \leq r \leq a, a \leq r \leq \sqrt{a^2+b^2}\right)$.

Now that we have established the result for $\int A_p$ we can start to establish the result for $\int C_p$. Given what we have so far we need to show that $\int C_p = 4r^3\left(\frac{R_1}{3} + \frac{R_2}{3} - \frac{r}{8}\right)$. Note that we only have the cases of p_6, p_7 and p_8 to consider because the connection radius of the other cases does not intersect with the inner rectangle.

6. Calculation of $\int_p C_p$ for points similar to p_6.

   We can start by investigating points similar to p_6. These points are such that the quarter circle crosses into the inner rectangle at a corner. To calculate this area we can follow the same procedure we used for calculating the red area from Figure 6.4 (b). Furthermore, if we let c denote the vertical distance a point is below the corner and h denote the horizontal distance it is to the left then this situation only occurs when $r < \sqrt{c^2+h^2}$, which is the exact same situation as for the summation of these areas for p_5. So we know that the contribution to $\int C_p$ from the corner segment is given by

$$\frac{r^4}{8} \tag{6.24}$$

   from Equation (6.21).

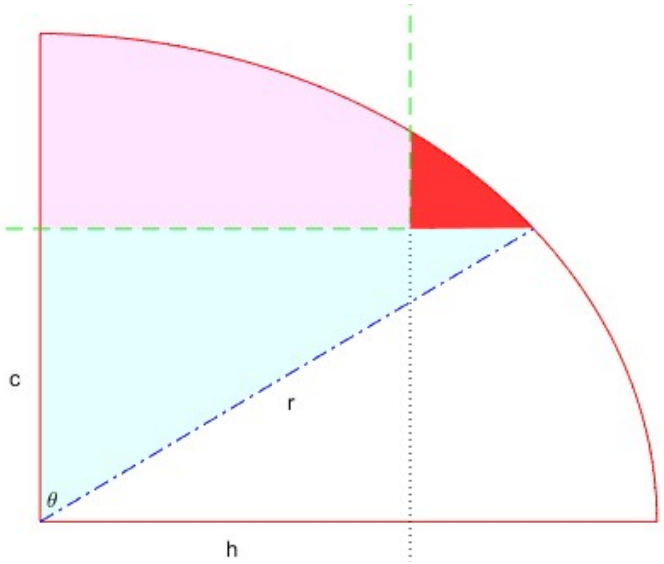


Figure 6.5: A depiction of the case for p_8 which can be worked out as the area for p_7 minus the red area which is back outside the inner rectangle.

7. Justification for treatment of points similar to p_8 in the same way as p_7.

   To assess the contribution to $\int_p C_p$ for points similar to p_7 we can consult Figure 6.3 (a) again. However, this time we are looking to assess the red area rather than the whole coloured area. The difference between a point similar to p_7 and a point similar to p_8 can be seen by looking at Figure 6.5. Essentially a point similar to p_8 can be calculated in the same way as a point similar to p_7 except that the dark red region in Figure 6.5 needs to be subtracted from the result. If we let c denote the vertical distance of a point from the bottom of the inner rectangle and h denote the horizontal distance it is to the left of its rightmost edge then points are similar to p_8 when $r < \sqrt{c^2 + h^2}$. This situation is once again the same as the one for the summation of the dark red region that we looked at when we considered the contribution to $\int_p A_p$ for points similar to p_5. That means that we already have the summation of such areas for all points similar to p_8 which was calculated in Equation (6.21) to be

$$\frac{r^4}{8}. \tag{6.25}$$

As a result we can treat points similar to p_8 the same as points similar to p_7 and work out the contribution across all such points and then subtract the result of Equation (6.25) at the end.

8. Calculation of $\int_p C_p$ for points similar to p_7 and p_8 which are below the inner rectangle.

   To work out the contribution to $\int_p C_p$ for points similar to p_7 we need to calculate the area of the red region in Figure 6.3 (a). We then integrate this result from 0 to r to get the sum of these results along a vertical line below the inner rectangle. Multiplying this result by R_1 will give the contribution for all points similar to p_7 and p_8 which are below the inner rectangle. The area of the red region in Figure 6.3 (a) is the area of the coloured segment with the area of the blue triangle subtracted which can be calculated by

$$\frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right) - \frac{c\sqrt{r^2-c^2}}{2}. \tag{6.26}$$

   We can integrate this area along the vertical line of length r perpendicular to the bottom of the rectangle using Equations (6.10) and (6.11)

$$\begin{aligned}&\int_0^r \frac{r^2}{2}\cos^{-1}\left(\frac{c}{r}\right) - \frac{c\sqrt{r^2-c^2}}{2}dc \\ &= \frac{r^3}{2} - \frac{r^3}{6} = \frac{r^3}{3}.\end{aligned} \tag{6.27}$$

   Multiplying this result by R_1 and subtracting the extra contribution for points similar to p_8 that we calculated in Equation (6.25) gives

$$\frac{R_1 r^3}{3} - \frac{r^4}{8}. \tag{6.28}$$

9. Calculation of $\int_p C_p$ for points similar to p_7 and p_8 which are to the left of the inner rectangle.

   Note that there are also some points which are to the left of the inner rectangle

and within r of it. We can work out the contribution of these points to $\int_p C_p$ analogously to the way we worked it out for the points below the inner rectangle. The only difference would be that we need to multiply the p_7 type points result by R_2 rather than R_1 which gives a contribution of such points as

$$\frac{R_2 r^3}{3} - \frac{r^4}{8}. \tag{6.29}$$

10. Final result of $\int_p C_p$.

    We can now calculate $\int C_p$ by combining Equations (6.24), (6.28) and (6.29) and multiplying the result by 4

$$\begin{aligned} \int C_p &= 4\left(\frac{R_1 r^3}{3} - \frac{r^4}{8} + \frac{R_2 r^3}{3} - \frac{r^4}{8} + \frac{r^4}{8}\right) \\ &= 4r^3\left(\frac{R_1}{3} + \frac{R_2}{3} - \frac{r}{8}\right). \end{aligned} \tag{6.30}$$

We are now in a position to put everything together and calculate

$$\mathbb{E}(\overline{k}) = \frac{(n-1)(\pi r^2(ab - R_1R_2) - \frac{4}{3}(a+b)r^3 + \frac{1}{2}r^4 - 4r^3\left(\frac{R_1}{3} + \frac{R_2}{3} - \frac{r}{8}\right)}{(ab - R_1R_2)^2}. \tag{6.31}$$

□

6.3 Effects of Number and Shape of Holes

We can generalise this situation to multiple holes provided that the holes are all at least the connection radius away from each other

Theorem 6.3. *Let there be a random rectangular annular graph with m holes all separated by a distance at least r and with $R_{1,j}, R_{2,j} > r$ for $1 \leq j \leq m$, where r is the connection radius. Then the expected average node degree of the graph is given by*

$$\begin{aligned}\mathbb{E}\left(\overline{k}\right) = &\frac{(n-1)\left(\pi r^2\left(ab-\Sigma_{j=1}^{m}\left(R_{1,j}R_{2,j}\right)\right)-\frac{4}{3}\left(a+b\right)r^3+\frac{1}{2}r^4\right)}{\left(ab-\Sigma_{j=1}^{m}\left(R_{1,j}R_{2,j}\right)\right)^2}\\ &-\frac{(n-1)\left(4r^3\left(\Sigma_{j=1}^{m}\left(\frac{R_{1,j}}{3}+\frac{R_{2,j}}{3}\right)-\frac{mr}{8}\right)\right)}{\left(ab-\Sigma_{j=1}^{m}\left(R_{1,j}R_{2,j}\right)\right)^2}.\end{aligned} \tag{6.32}$$

Proof. Given that the boundary of each hole is at least r from the boundary of any other hole then each extra hole does not alter the boundary effects of any hole which was already present and we can safely sum the boundary effects of each individual hole to derive the total effect. □

We are now in a position where we can prove some results regarding the expected average node degree and the number size and shape of areas in which nodes cannot be scattered on random rectangular annular graphs.

Theorem 6.4. *Let a, b be fixed side lengths of the outer boundary of a random rectangular annular graph, G_c, with one annular area with side lengths $R_1 = Rc, R_2 = \frac{R}{c}$ where $c > 1$. Then if $d > c$ then $\mathbb{E}(\overline{k})_{G_c} > \mathbb{E}(\overline{k})_{G_d}$.*

Proof. We can deduce that $\mathbb{E}(\overline{k})_{G_c} > \mathbb{E}(\overline{k})_{G_d}$ if and only if $d - c > \frac{(d-c)}{cd}$ using the following series of equivalences

$$\begin{aligned}&\mathbb{E}(\overline{k})_{G_c} > \mathbb{E}(\overline{k})_{G_d}\\ \iff &\frac{\left(\pi r^2(ab-\frac{RcR}{c})-4r^3\left(\frac{Rc}{3}+\frac{R}{3c}-\frac{r}{8}\right)\right)}{\left(ab-\frac{RcR}{c}\right)^2} > \frac{\left(\pi r^2(ab-\frac{RdR}{d})-4r^3\left(\frac{Rd}{3}+\frac{R}{3d}-\frac{r}{8}\right)\right)}{\left(ab-\frac{RdR}{d}\right)^2}\\ \iff &-4r^3\left(\frac{Rc}{3}+\frac{R}{3c}-\frac{r}{8}\right) > -4r^3\left(\frac{Rd}{3}+\frac{R}{3d}-\frac{r}{8}\right)\\ \iff &\frac{Rd}{3}+\frac{R}{3d} > \frac{Rc}{3}+\frac{R}{3c}\\ \iff &d+\frac{1}{d} > c+\frac{1}{c}\\ \iff &d-c > \frac{(d-c)}{cd}.\end{aligned} \tag{6.33}$$

By definition $c, d > 1$ and the inequality holds. □

It makes sense that a more elongated zone would cause a greater decrease in the expected average node centrality than one which had a shape closer to a square. A more elongated shape requires a larger perimeter to cover the same area and the larger the perimeter of the zone in which nodes cannot be spread the larger the area within the connection radius r of said zone. So more nodes would be expected to fall into an area where the circle with radius r centred at their location would intersect with the zone where nodes cannot be spread and such nodes are expected to have lower node degrees.

The same is true if we were to split an area into two pieces.

Theorem 6.5. *Let G_1 be a random rectangular annular graph with a hole with side lengths R_1, R_2. Let G_2 be a second Random Rectangular Annular Graph with two holes which have side lengths cR_1, R_2 and dR_1, R_2 respectively where $c+d=1$. Let a, b be the side lengths of the outer boundary for both G_1 and G_2. Then $\mathbb{E}(\overline{k})_{G_1} > \mathbb{E}(\overline{k})_{G_2}$.*

Proof. We need to show that $\mathbb{E}(\overline{k})_{G_1} > \mathbb{E}(\overline{k})_{G_2}$. We start by noting that

$$R_1R_2 = (c+d)R_1R_2 = cR_1R_2 + dR_1R_2. \tag{6.34}$$

We can demonstrate that the theorem is true if and only if $\frac{r}{8} < \frac{R_2}{3}$ using the following series of inequalities.

$$\begin{aligned}
&\mathbb{E}(\overline{k})_{G_1} > \mathbb{E}(\overline{k})_{G_2} \\
\iff &-4r^3\left(\frac{R_1}{3} + \frac{R_2}{3} - \frac{r}{8}\right) > -4r^3\left(\frac{cR_1}{3} + \frac{R_2}{3} + \frac{dR_1}{3} + \frac{R_2}{3} - \frac{2r}{8}\right) \\
\iff &-4r^3\left(\frac{R_1}{3} + \frac{R_2}{3} - \frac{r}{8}\right) > -4r^3\left(\frac{R_1}{3} + \frac{2R_2}{3} - \frac{2r}{8}\right) \\
\iff &\frac{R_1}{3} + \frac{R_2}{3} - \frac{r}{8} < \frac{R_1}{3} + \frac{2R_2}{3} - \frac{2r}{8} \\
\iff &\frac{r}{8} < \frac{R_2}{3}
\end{aligned} \tag{6.35}$$

We have that $R_2 > r$ by definition and the inequality holds. □

Again this theorem makes sense because of the greater perimeter of the annular zones of G_2 than G_1, there being four sides of length R_2 compared to two. Therefore, a larger portion of nodes are expected to be within r of an annular zone.

For the same reason, it is also true that two smaller rectangular annular zones cause a greater decrease in $\mathbb{E}(\overline{k})$ than one large one covering the same total area, provided that the proportions of the sides are the same for all three rectangles.

Theorem 6.6. *Let G_1 be a random rectangular annular graph with a hole with side lengths R_1, R_2. Let G_2 be a second Random Rectangular Annular Graph with two holes which have side lengths cR_1, cR_2 and dR_1, dR_2 respectively where $c^2 + d^2 = 1$. Let a, b be the side lengths of the outer boundary for G_1 and G_2. Then we have $\mathbb{E}(\overline{k})_{G_1} \geq \mathbb{E}(\overline{k})_{G_2}$.*

Proof. Without loss of generality we can set $c \geq d$ and $R_1 \geq R_2$. It should also be noted that by the definition of a random rectangular annular graph $r < R_2 d$. We note that

$$(c^2 R_1 R_2 + d^2 R_1 R_2) = (c^2 + d^2) R_1 R_2 = R_1 R_2. \tag{6.36}$$

We show that if $d(416 - 425d) \geq 0$ then $\mathbb{E}(\overline{k})_{G_1} \geq \mathbb{E}(\overline{k})_{G_2}$. It is the case that $d(416 - 425d) \geq 0$ because $d \leq c$ which means that $d \leq \frac{1}{\sqrt{2}}$ hence $d \leq \frac{416}{425}$ and clearly $d \geq 0$. We start by deriving a reduced inequality for $\mathbb{E}(\overline{k})_{G_1} \geq \mathbb{E}(\overline{k})_{G_2}$.

$$\begin{aligned}
& \mathbb{E}(\overline{k})_{G_1} \geq \mathbb{E}(\overline{k})_{G_2} \\
\iff \quad & -4r^3 \left(\frac{R_1}{3} + \frac{R_2}{3} - \frac{r}{8} \right) \geq -4r^3 \left(\frac{cR_1}{3} + \frac{cR_2}{3} + \frac{dR_1}{3} + \frac{dR_2}{3} - \frac{2r}{8} \right) \\
\iff \quad & \frac{(c+d)R_1}{3} + \frac{(c+d)R_2}{3} - \frac{r}{8} \geq \frac{R_1}{3} + \frac{R_2}{3} \\
\iff \quad & \frac{(c+d-1)R_1}{3} + \frac{(c+d-1)R_2}{3} \geq \frac{r}{8}
\end{aligned} \tag{6.37}$$

We have from Equation (6.37) that $\mathbb{E}(\overline{k})_{G_1} \geq \mathbb{E}(\overline{k})_{G_2}$ if and only if $\frac{(c+d-1)R_1}{3} + \frac{(c+d-1)R_2}{3} \geq \frac{r}{8}$. We also have that $R_1 \geq R_2$ and $r \leq R_2 d$ which means that if $\frac{2(c+d-1)R_2}{3} \geq \frac{R_2 d}{8}$ then $\frac{(c+d-1)R_1}{3} + \frac{(c+d-1)R_2}{3} \geq \frac{r}{8}$. We show that $\frac{2(c+d-1)R_2}{3} \geq \frac{R_2 d}{8}$ if and only if $d(416 - 425d) \geq 0$.

$$
\begin{aligned}
& \frac{2(c+d-1)R_2}{3} \geq \frac{R_2 d}{8} \\
\iff \quad & 16(\sqrt{1-d^2}+d-1)R_2 \geq 3R_2 d \\
\iff \quad & R_2(16\sqrt{1-d^2}+13d-16) \geq 0 \\
\iff \quad & 16\sqrt{1-d^2} \geq 16-13d \\
\iff \quad & 256(1-d^2) \geq 256-416d+169d^2 \\
\iff \quad & 416d-425d^2 \geq 0 \\
\iff \quad & d(416-425d) \geq 0
\end{aligned}
\tag{6.38}
$$

□

It should be noted that it is not always the case that two holes will cause a greater decrease in the expected average degree centrality of a random rectangular annular graph than one hole. For instance a single zone covering an area of 1 by 8 will cause a decrease of $4r^3\left(\frac{1}{3}+\frac{8}{3}-\frac{r}{3}\right)$ which is larger than the decrease caused by two square zones of side length 2, $4r^3\left(\frac{2}{3}+\frac{2}{3}+\frac{2}{3}+\frac{2}{3}-\frac{2r}{3}\right)$. Again we can see the length of the perimeter coming into effect here. The total length of the perimeter of the 8 by 1 area is 18 whereas the combined perimeter of the two side length 2 squares is 16.

However, the general pattern is clear. Multiple small areas, on which nodes cannot be spread, within the rectangle will tend to cause a larger decrease in the expected average degree centrality of the resulting network than a few larger areas where the total area covered in the two cases is the same. Where the proportions of the side lengths of these areas is fixed or the larger areas could be formed by amalgamating smaller ones then we have shown that this difference occurs.

6.4 SIS/SIR Models on RRAGs

In Section 2.4 we discussed models of disease spread on networks. These models could be used to model the systems of plant disease which we discussed earlier in this chapter and which has already been done by Estrada et al. [47] and Arias et al. [7]. In Section 2.4 we also presented an approximate proof of Theorem 2.24 which was proved without the assumptions used here by Van Miegham et al. [166]. This theorem connected

the epidemic threshold of a network to its spectral radius through the relationship $\tau = \frac{\beta}{\gamma} = \frac{1}{\lambda_1}$. We also have from Theorem 2.14 that $\lambda_1 \geq \overline{k}$ which means that we have a bound on the epidemic threshold of a network

$$\frac{1}{\lambda_1} \leq \frac{1}{\overline{k}}. \tag{6.39}$$

In the previous section we demonstrated that increasing the number of holes resulted in a decrease in the expected average degree, $\mathbb{E}(\overline{k})$, in a random rectangular annular graph, provided that the same area is covered and that the holes are of the same shape. Then it is not unreasonable to suspect that, subject to the same constraints, increasing the number of holes will also result in an increase in the epidemic threshold on a random rectangular annular graph.

6.5 Simulations

To investigate the effect of the number of holes on random rectangular annular graphs we have focused on 6 cases and also included a control on a regular RGG. In all 6 cases the simulations were done on RRAGs which had 1000 nodes, and a connection radius of $r = 0.07$. The outer square has side length 1 in all cases and the total area on which nodes cannot be scattered is 0.25, representing $\frac{1}{4}$ of the total area. The 7^{th} case is a control which does not have a hole. It has the same connection radius as the 6 cases under study but has 1333 nodes to maintain the same node density as them. All simulations are based on 20 realisations with the largest eigenvalue and the mean degree calculated for each case.

In the first four cases, which are shown in Figure 6.6, we look into RRAGs with different numbers and sizes of holes. The cases chosen are a single central square hole of side length $\frac{1}{2}$, 4 square holes of side length $\frac{1}{4}$, 9 square holes of side length $\frac{1}{6}$ and 16 holes of side length $\frac{1}{8}$. These choices will allow us to determine the effect that having a single large hole may have on the epidemic threshold versus many small holes. We would expect from the calculations of the expected average degree that many small holes would cause an increase to the epidemic threshold.

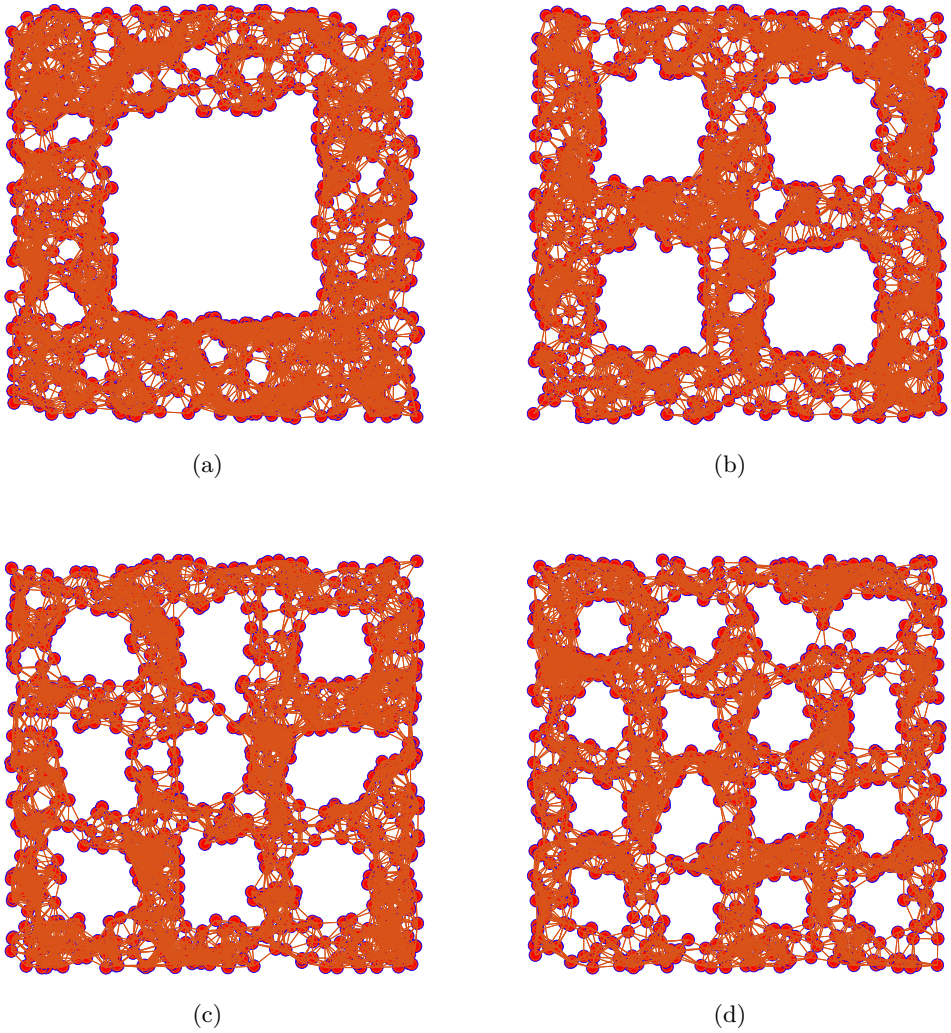


Figure 6.6: Example of an RGG with a single central square hole of side length $\frac{1}{2}$ (a), 4 square holes of side length $\frac{1}{4}$ (b), 9 square holes of side length $\frac{1}{6}$ (c) and 16 square holes of side length $\frac{1}{8}$ (d). In each case the centre points of the holes are equidistant from each other and from the boundary of the outer square. All RGGs were done using 1000 nodes, with a connection radius of $r = 0.07$ with the outer square having side length 1 and a total missing area of 0.25.

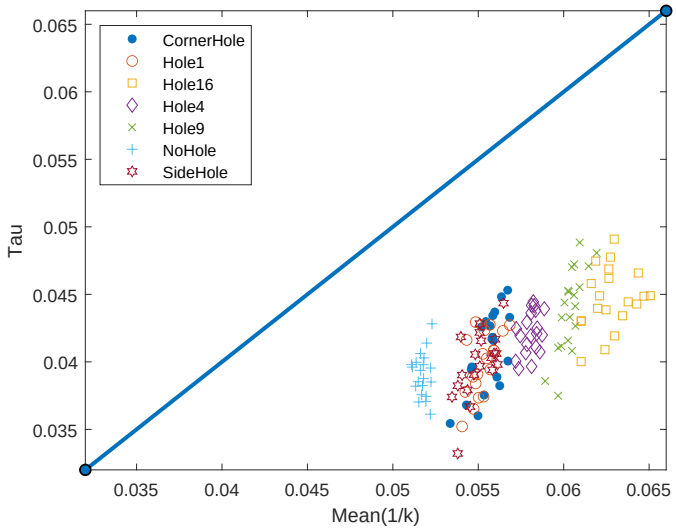


Figure 6.7: The results for the individual simulations with the reciprocal of the mean degree on the x-axis and the epidemic threshold on the y-axis. The identity line is also shown to demonstrate the relationship between change in the epidemic threshold and the change in the reciprocal of the mean degree.

The 5^{th} and 6^{th} cases investigate whether moving the location of the hole within the RRAG affects the dynamics. In the 5^{th} case the large central hole of the RRAG in Figure 6.6 (a) is moved to $(0.32, 0.5)$ so that its centre point is closer to the edge of the zone and in the 6^{th} scenario it is as close as possible to the corner $(0.32, 0.32)$. A comparison between these cases will allow us to assess whether the location of the hole has an effect on the epidemic threshold. We expect that there would not be a difference between the two cases given that the expected mean degree was not dependent on the location of the hole.

The results of these simulations are shown in Figure 6.7 with mean results across the 20 cases displayed in Table 6.1.

From Table 6.1 and Figure 6.7 we can see that there does seem to be an effect in terms of an increased epidemic threshold when the obstacles which are in the way of transmission are many and small as opposed to single large obstacles. To assess the statistical significance of this trend we applied the Mann-Whitney U test, which we introduced in Section 2.7.3, to each pairing of cases to detect whether the difference between the distribution of the epidemic thresholds in each pair is significant at the 5%

Case	$Median(\tau)$	$\overline{\tau}$	$\overline{(\overline{k})}$	$E(\overline{k})$	$\overline{(\frac{1}{k})}$
Central Hole	0.0404	0.0400	18.1053	18.1105	0.0552
4 holes	0.0420	0.0421	17.2216	17.3623	0.0581
9 holes	0.0443	0.0439	16.5583	16.6566	0.0604
16 holes	0.0447	0.0447	15.9401	15.9937	0.0628
Side Hole	0.0401	0.0398	18.1837	18.1105	0.0550
Corner Hole	0.0416	0.0408	17.9924	18.1105	0.0556
Control	0.0391	0.0390	19.3284	19.3022	0.0517

Table 6.1: This table details the mean epidemic threshold and the mean of the mean degree across the 20 runs of each case. The expected mean degree of each case is also displayed to verify the validity of the formula derived in Section 6.2.

Case	4 holes	9 holes	16 holes	Edge	Corner	Control
Central Hole	0.0004	<0.0001	<0.0001	0.7533	0.1732	0.0642
4 holes		0.0052	<0.0001	<0.0001	0.0694	<0.0001
9 holes			0.4423	<0.0001	0.0003	<0.0001
16 holes				<0.0001	<0.0001	<0.0001
Side Hole					0.1522	0.0361
Corner Hole						0.0080

Table 6.2: This table displays the p-value resulting from a Mann-Whitney U test between the distributions of the 20 epidemic thresholds for each pair of cases tested.

level. Given that there are 21 pairings there is an increased chance that at least one of the conclusions drawn is erroneous compared to if we were only doing one comparison. The p-value generated by each pairing is displayed in Table 6.2.

The Mann-Whitney U test results largely also support the conclusions that we have drawn regarding there being evidence of a change in the epidemic threshold when there are several small holes in comparison to one large one. The general trend is also that there is no evidence to suggest that hole positioning influences the epidemic threshold. There are, however, exceptions. The lack of evidence of a difference between the 9 hole case and the 16 hole case suggests that there is perhaps a law of diminishing returns in terms of splitting the area into more small areas. The other two results which disagree with what we had put forward are the lack of evidence of a difference between the control and the RRAG with one large hole in the middle and similarly between the 4 holes case and the 1 hole in the corner case despite the evidence of a difference between the 4 holes case and the other two 1 hole cases.

Chapter 7

Conclusion

7.1 Results

We have generalised several concepts from network theory into the realm of simplicial complexes, principally concepts of degree distribution and centrality measures. We have then applied these concepts to protein–protein interaction networks to demonstrate that, at least for the interactomes studied, the degree distributions at the level of the 1 and 2-simplices feature many simplices of medium degree which differs from the situation for nodes where there are a lot of nodes of low degree and only a small number of medium and high degree ones. The other key application in terms of the PPI networks was the use of simplicial centrality measures to increase the number of essential proteins it was possible to find in the yeast cell compared to their node-based equivalents.

We also discussed another use case for the simplicial centralities in terms of wireless sensor networks where we demonstrated that the edge centralities for simplicial complexes were likely to identify edges and nodes which were relatively isolated but close to dense areas. This finding means that it is possible to open up gaps in coverage near quite well covered areas by selecting just a few key nodes to remove which is useful to both those actors who would like to disrupt communication in systems which can be modelled by random geometric graphs and those actors who wish to identify areas which are vulnerable to such attacks or random failures and make their systems more

robust.

The main takeaway from both of these sets of results should be that the use of simplicial complexes opens up information which would not be available if our analyses were restricted to only the previously available network-theoretic measures. It is likely that there are many other areas of investigation which would benefit from being looked at under a simplicial complexes framework. Although care should clearly be taken to ensure that any such application is justified from a physical or sociological perspective before it is undertaken.

The final key takeaway from this work is the derivation of an equation for the expected mean degree of a random rectangular annular graph where there are obstacles on which it is not possible to distribute nodes. We demonstrated several cases by which it was possible to decrease this average degree including elongation of these areas, splitting these areas up and constructing many smaller such areas of the same shape. We then connected this result to the phenomenon of epidemic spreading and produced evidence that the epidemic threshold is expected to increase more in the presence of many small holes rather than a few larger ones. We also showed that neither the mean degree or the epidemic threshold were affected by the positioning of these holes.

7.2 Suggestions for Future Work

One key improvement of the work undertaken here would be to use more reliable protein–protein interaction data to fully leverage the power of simplicial complexes in detecting essential proteins if it becomes possible to separate the case of multiple pairwise interactions from that of one interaction with multiple actors.

Another interesting concept which the author did not get to would be to study consensus dynamics on networks with holes. This problem is difficult because hole positioning matters when it comes to the consensus dynamics which makes investigating the effects of multiple holes a lot more difficult that it was for the epidemic spreading case. However, it should be interesting nonetheless.

We also think it would be interesting to see how theoretical and simulated results of epidemic spreading on random geometric graphs compare to a real situation. The case

of the spread of invasive plants comes to mind as a potential application. Rhododendron ponticum propagates itself using the wind which can carry it for up to 100m. It also needs particular conditions in which to grow. It is particularly invasive on the west coast of Scotland, a region marked by mountainous areas and lochs which would be barriers to its spread.

The final suggestion does not relate directly to the work undertaken in this thesis. There is work being done on the flapper skate populations off Scotland's west coast which is attempting to identify individual fish to get a better estimate of the population. These fish have spot patterns on their backs. It may be possible to leverage the multiplicative persistence of holes introduced by Bobrowski et al. [20] to compare these spot patterns as each individual fish grows and create a unique and scalable barcode for each fish. Variations of this multiplicative persistence would also need to be introduced in order to handle photographs being taken from different angles. Access to these tools would allow researchers to semi-automate the identification of individual fish which is currently done by judging photographs by eye. The use of spot pattern analysis to identify individuals within a species is a common technique [103, 115] which has been the subject of automation [28] and if the technique outlined above proves effective it could be generalised to other species.

List of Symbols

λ_1	Largest eigenvalue of a matrix
B^T	Transpose of a matrix B
B^{-1}	Inverse of a matrix B
I_n	Identity matrix of size n
δ_{ij}	Kronecker delta
δ_k	degree of a k-simplex
$l(m, c)$	The line with gradient m and intercept c
$Rot_\theta(v)$	Rotation of vector v by θ with respect to the origin
$T_{a,b}(v)$	Translation of a vector v by a in the x-coordinate and b in the y-coordinate
$Ref_y(v)$	Reflection of a vector v through the y-axis
k_p	Degree of a node p in a network
$\overline{k}$	Average degree across a network
$\tau = \frac{1}{\lambda}$	The epidemic threshold of a network
L_k	Hodge Laplacian for the k-simplices
A_k	Simplex Adjacency Matrix for the k-simplices
∂_k	Boundary operator on the k-chains
k-	Signifies a network-theoretic concept on the k-simplices
C_k	The group of k-chains
Z_k	The kernel of ∂_k
B_k	The image of ∂_{k+1}
H_k	The k^{th} Homology group of a simplicial complex
β_k	The k^{th} Betti number of a simplicial complex

rp	The Pearson Correlation coefficient
rs	The Spearman Correlation coefficient
l_k	The average k-simplicial shortest path distance for the k-simplices
CC_k	Closeness centrality on the k-simplices
HC_k	Harmonic Closeness centrality on the k-simplices
SC_k	Subgraph centrality on the k-simplices
d_k	Shortest-path distance on the k-simplices
r	Connection radius

Bibliography

[1] V. Agarwal, S. Tapaswi, and P. Chanak. A Survey on Path Planning Techniques for Mobile Sink in IoT-Enabled Wireless Sensor Networks. *Wireless Personal Communications*, 119(1):211–238, 2021.

[2] A. Allen-Perkins. Random Spherical Graphs. *Physical Review E*, 98(3):032310, 2018.

[3] L. Alonso, J. A. Méndez-Bermúdez, A. Gonzalez-Melendrez, and Y. Moreno. Weighted Random-Geometric and Random-Rectangular Graphs: Spectral and Eigenfunction Properties of the Adjacency Matrix. *Journal of Complex Networks*, 6(5):753–766, 2018.

[4] D. V. Anand and M. K. Chung. Hodge Laplacian of Brain Networks. *IEEE Transactions on Medical Imaging*, 42(5):1563–1573, 2023.

[5] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[6] O. Angel and Y. Spinka. Geometric Random Graphs on Circles. In *Advances in Probability and Mathematical Statistics: CLAPEM 2019, Mérida, Mexico*, pages 23–41. Springer, 2021.

[7] J. H. Arias, J. Gómez-Gardeñes, S. Meloni, and E. Estrada. Epidemics on Plants: Modeling Long-Range Dispersal on Spatially Embedded Networks. *Journal of Theoretical Biology*, 453:1–13, 2018.

[8] F. Arrigo, D. J. Higham, and F. Tudisco. A Framework for Second-Order Eigenvector Centralities and Clustering Coefficients. *Proceedings of the Royal Society A*, 476(2236):20190724, 2020.

[9] R. H. Atkin. From Cohomology in Physics to Q-Connectivity in Social Science. *International Journal of Man-Machine Studies*, 4(2):139–167, 1972.

[10] S. Bagali and R. Sundaraguru. Efficient Resource Allocation Design for Mitigating Multi-Jammer in Underwater Wireless Sensor Network. *EAI Endorsed Transactions on Internet of Things*, 6(24):e2–e2, 2021.

[11] N. T. Bailey et al. *The Mathematical Theory of Infectious Diseases and its Applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[12] D. S. Bassett, P. Zurn, and J. I. Gold. On the Nature and Use of Models in Network Neuroscience. *Nature Reviews Neuroscience*, 19(9):566–578, 2018.

[13] F. Baudoin, S. Rigot, G. Savaré, and N. Shanmugalingam. *New Trends on Analysis and Geometry in Metric Spaces*. Springer, 2022.

[14] A. Bavelas. Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

[15] A. R. Benson. Three Hypergraph Eigenvector Centralities. *SIAM Journal on Mathematics of Data Science*, 1(2):293–312, 2019.

[16] G. Bianconi. *Higher-Order Networks*. Cambridge University Press, 2021.

[17] G. Bianconi and C. Rahmede. Complex Quantum Network Manifolds in Dimension d > 2 are Scale-Free. *Scientific Reports*, 5(1):13979, 2015.

[18] N. Biggs. *Discrete Mathematics*. Oxford University Press, 2002.

[19] D. B. Blumenthal, M. Lucchetta, L. Kleist, S. P. Fekete, M. List, and M. H. Schaefer. Emergence of Power Law Distributions in Protein-Protein Interaction Networks Through Study Bias. *eLife*, 13:e99951, 2024.

[20] O. Bobrowski, M. Kahle, and P. Skraba. Maximally Persistent Cycles in Random Geometric Complexes. *The Annals of Applied Probability*, pages 2032–2060, 2017.

[21] A. Bravais. Mathematical Analysis on the Probabilities of the Situation Errors of a Point. *Memories Presented by Various Scientists. The Royal Academy of Sciences of the Institute de France*, 9:255–332, 1846.

[22] A. Bretto. Hypergraph Theory. *An Introduction. Mathematical Engineering. Cham: Springer*, 1:209–216, 2013.

[23] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[24] S. Brohee and J. Van Helden. Evaluation of Clustering Algorithms for Protein–Protein Interaction Networks. *BMC Bioinformatics*, 7(1):1–19, 2006.

[25] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, et al. Topological Structure Analysis of the Protein–Protein Interaction Network in Budding Yeast. *Nucleic Acids Research*, 31(9):2443–2450, 2003.

[26] G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, et al. Interaction Network Containing Conserved and Essential Protein Complexes in Escherichia coli. *Nature*, 433(7025):531–537, 2005.

[27] P. Chanak, I. Banerjee, and R. S. Sherratt. Energy-Aware Distributed Routing Algorithm to Tolerate Network Failure in Wireless Sensor Networks. *Ad Hoc Networks*, 56:158–172, 2017.

[28] G. S. Cheema and S. Anand. Automatic Detection and Recognition of Individuals in Patterned Species. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III 10*, pages 27–38. Springer, 2017.

[29] S. Chen, C. Huang, L. Wang, and S. Zhou. A Disease-Related Essential Protein Prediction Model Based on the Transfer Neural Network. *Frontiers in Genetics*, 13:1087294, 2023.

[30] H. Chintakunta and H. Krim. Distributed Localization of Coverage Holes Using Topological Persistence. *IEEE Transactions on Signal Processing*, 62(10):2531–2541, 2014.

[31] F. R. Chung. *Spectral Graph Theory*, volume 92. American Mathematical Society, 1997.

[32] L. E. Da Rocha. Structural Evolution of the Brazilian Airport Network. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(04):P04020, 2009.

[33] H.-L. De Kergorlay and D. J. Higham. Connectivity of Random Geometric Hypergraphs. *Entropy*, 25(11):1555, 2023.

[34] V. De Silva and R. Ghrist. Coverage in Sensor Networks Via Persistent Homology. *Algebraic & Geometric Topology*, 7(1):339–358, 2007.

[35] J. N. Deshpande. *Dynamiques Éco-Évolutives: des Réseaux de Gènes aux Paysages Complexes*. PhD thesis, Université de Montpellier, 2024.

[36] A. Duncan. Powers of the Adjacency Matrix and the Walk Matrix. *The Collection*, 9:1–11, 2004.

[37] J. A. Dunne. The Network Structure of Food Webs. In *Ecological Networks: Linking Structure to Dynamics in Food Webs*, pages 27–86, 2006.

[38] B. Eckmann. Harmonische Funktionen und Randwertaufgaben in einem Komplex. *Commentarii Mathematici Helvetici*, 17(1):240–255, 1944.

[39] H. Edelsbrunner. *A Short Course in Computational Geometry and Topology*. Springer, 2014.

[40] L. Elcoro, Z. Song, and B. A. Bernevig. Application of Induction Procedure and Smith Decomposition in Calculation and Topological Classification of Electronic

Band Structures in the 230 Space Groups. *Physical Review B*, 102(3):035110, 2020.

[41] P. Erdős and A. Rényi. On the Evolution of Random Graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60, 1960.

[42] E. Estrada. Virtual Identification of Essential Proteins within the Protein Interaction Network of Yeast. *Proteomics*, 6(1):35–40, 2006.

[43] E. Estrada. *The Structure of Complex Networks: Theory and Applications*. American Chemical Society, 2012.

[44] E. Estrada and Ö. Bodin. Using Network Centrality Measures to Manage Landscape Connectivity. *Ecological Applications*, 18(7):1810–1825, 2008.

[45] E. Estrada and G. Chen. Synchronizability of Random Rectangular Graphs. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(8):083107, 2015.

[46] E. Estrada and P. A. Knight. *A First Course in Network Theory*. Oxford University Press, New York, NY, USA, 2015.

[47] E. Estrada, S. Meloni, M. Sheerin, and Y. Moreno. Epidemic Spreading in Random Rectangular Networks. *Physical Review E*, 94(5):052316, 2016.

[48] E. Estrada and J. A. Rodriguez-Velazquez. Subgraph Centrality in Complex Networks. *Physical Review E*, 71(5):056103, 2005.

[49] E. Estrada and J. A. Rodríguez-Velázquez. Subgraph Centrality and Clustering in Complex Hyper-Networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594, 2006.

[50] E. Estrada and G. J. Ross. Centralities in Simplicial Complexes. Applications to Protein Interaction Networks. *Journal of Theoretical Biology*, 438:46–60, 2018.

[51] E. Estrada and M. Sheerin. Random Rectangular Graphs. *Physical Review E*, 91(4):042805, 2015.

[52] E. Estrada and M. Sheerin. Consensus Dynamics on Random Rectangular Graphs. *Physica D: Nonlinear Phenomena*, 323:20–26, 2016.

[53] L. Euler. Solutio Problematis Ad Geometriam Situs Pertinentis. *Commentarii Academiae Scientiarum Petropolitanae*, pages 128–140, 1741.

[54] L. Euler. The Seven Bridges of Königsberg. *The World of Mathematics*, 1:573–580, 1956.

[55] M. G. Everett and S. P. Borgatti. Extending Centrality. *Models and Methods in Social Network Analysis*, 35(1):57–76, 2005.

[56] A. Flammini, P. Ferrari, D. Marioli, E. Sisinni, and A. Taroni. Wired and Wireless Sensor Networks for Industrial Applications. *Microelectronics Journal*, 40(9):1322–1336, 2009.

[57] L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40:35–41, 1977.

[58] S. Galhotra, A. Mazumdar, S. Pal, and B. Saha. Connectivity in Random Annulus Graphs and the Geometric Block Model. *arXiv preprint arXiv:1804.05013*, 2018.

[59] H. Ge, Z. Liu, G. M. Church, and M. Vidal. Correlation Between Transcriptome and Interactome Mapping Data from Saccharomyces cerevisiae. *Nature Genetics*, 29(4):482–486, 2001.

[60] Y. Y. Ghadi, T. Mazhar, T. Al Shloul, T. Shahzad, U. A. Salaria, A. Ahmed, and H. Hamam. Machine Learning Solutions for the Security of Wireless Sensor Networks: A Review. *IEEE Access*, 12:12699–12719, 2024.

[61] K. Ghoumid, D. Ar-Reyouchi, S. Rattal, R. Yahiaoui, O. Elmazria, et al. Protocol Wireless Medical Sensor Networks in IoT for the Efficiency of Healthcare. *IEEE Internet of Things Journal*, 9(13):10693–10704, 2021.

[62] R. W. Ghrist. *Elementary Applied Topology*, volume 1. Createspace Seattle, 2014.

[63] E. N. Gilbert. Random Plane Networks. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):533–543, 1961.

[64] A. P. Giles, O. Georgiou, and C. P. Dettmann. Connectivity of Soft Random Geometric Graphs Over Annuli. *Journal of Statistical Physics*, 162:1068–1083, 2016.

[65] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. Hao, C. Ooi, B. Godwin, E. Vitols, et al. A Protein Interaction Map of Drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

[66] T. E. Goldberg. Combinatorial Laplacians of Simplicial Complexes. *Senior Thesis, Bard College*, 2002.

[67] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno. Discrete-Time Markov Chain Approach to Contact-Based Disease Spreading in Complex Networks. *Europhysics Letters*, 89(3):38009, 2010.

[68] R. Guimerà, S. Mossa, A. Turtschi, and L. N. Amaral. The Worldwide Air Transportation Network: Anomalous Centrality, Community Structure, and Cities' Global Roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, 2005.

[69] A. M. Gustafson, E. S. Snitkin, S. C. Parker, C. DeLisi, and S. Kasif. Towards the Identification of Essential Genes Using Targeted Genome Sequencing and Comparative Analysis. *BMC Genomics*, 7(1):265, 2006.

[70] D. Horak and J. Jost. Spectra of Combinatorial Laplace Operators on Simplicial Complexes. *Advances in Mathematics*, 244:303–336, 2013.

[71] D. Horak, S. Maletić, and M. Rajković. Persistent Homology of Complex Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03034, 2009.

[72] A. Jenkins. Community Detection via Persistent Homology Surfaces. Master's thesis, Brigham Young University, 2023.

[73] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and Centrality in Protein Networks. *Nature*, 411(6833):41–42, 2001.

[74] B. Jiang and I. Omer. Spatial Topology and its Structural Analysis Based on the Concept of Simplicial Complex. *Transactions in GIS*, 11(6):943–960, 2007.

[75] Y. Jiang, Y. Wang, W. Pang, L. Chen, H. Sun, Y. Liang, and E. Blanzieri. Essential Protein Identification Based on Essential Protein–Protein Interaction Prediction by Integrated Edge Weights. *Methods*, 83:51–62, 2015.

[76] S. Jin, Z. Huang, Y. Chen, D. Chavarría-Miranda, J. Feo, and P. C. Wong. A Novel Application of Parallel Betweenness Centrality to Power Grid Contingency Analysis. In *2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS)*, pages 1–7. IEEE, 2010.

[77] M. Kahle. Random Geometric Complexes. *Discrete & Computational Geometry*, 45:553–573, 2011.

[78] M. Kahle. Random Simplicial Complexes. In *Handbook of Discrete and Computational Geometry*, pages 581–603. Chapman and Hall/CRC, 2017.

[79] I. Kapovich and R. Weidmann. Kleinian Groups and the Rank Problem. *Geometry & Topology*, 9(1):375–402, 2005.

[80] R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[81] L. Katz. A New Status Index Derived from Sociometric Analysis. *Psychometrika*, 18(1):39–43, 1953.

[82] H. Kenniche and V. Ravelomananana. Random Geometric Graphs as Model of Wireless Sensor Networks. In *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, volume 4, pages 103–107. IEEE, 2010.

[83] H. Khojasteh, A. Khanteymoori, and M. H. Olyaee. Comparing Protein–Protein Interaction Networks of SARS-CoV-2 and (H1N1) Influenza using Topological Features. *Scientific Reports*, 12(1):5867, 2022.

[84] T. K. Kim and J. H. Park. More About the Basic Assumptions of t-test: Normality and Sample Size. *Korean Journal of Anesthesiology*, 72(4):331–335, 2019.

[85] A. P. King and R. Eckersley. *Statistics for Biomedical Engineers and Scientists: How to Visualize and Analyze Data*. Academic Press, 2019.

[86] O. Kirk, T. V. Borchert, and C. C. Fuglsang. Industrial Enzyme Applications. *Current Opinion in Biotechnology*, 13(4):345–351, 2002.

[87] F. Klimm, C. M. Deane, and G. Reinert. Hypergraphs for Predicting Essential Genes Using Multiprotein Complex Data. *Journal of Complex Networks*, 9(2):cnaa028, 2021.

[88] S. Konishi and G. Kitagawa. *Information Criteria and Statistical Modeling*. Springer Science & Business Media, New York, NY, USA, 2008.

[89] V. E. Krebs. Mapping Networks of Terrorist Cells. *Connections*, 24(3):43–52, 2002.

[90] D. J. LaCount, M. Vignali, R. Chettier, A. Phansalkar, et al. A Protein Interaction Network of the Malaria Parasite Plasmodium falciparum. *Nature*, 438(7064):103, 2005.

[91] N. Lasla, M. F. Younis, A. Ouadjaout, and N. Badache. An Effective Area-Based Localization Algorithm for Wireless Networks. *IEEE Transactions on Computers*, 64(8):2103–2118, 2014.

[92] S. Lawson, D. Donovan, and J. Lefevre. An Application of Node and Edge Nonlinear Hypergraph Centrality to a Protein Complex Hypernetwork. *PLOS One*, 19(10):e0311433, 2024.

[93] M. Li, H. Zhang, J.-X. Wang, and Y. Pan. A New Essential Protein Discovery Method Based on the Integration of Protein–Protein Interaction and Gene Expression Data. *BMC Systems Biology*, 6:1–9, 2012.

[94] C.-Y. Lin, C.-L. Chen, C.-S. Cho, L.-M. Wang, C.-M. Chang, P.-Y. Chen, C.-Z. Lo, and C. A. Hsiung. HP-DPI: Helicobacter pylori Database of Protein Interactomes Embracing-Experimental and Inferred Interactions. *Bioinformatics*, 21(7):1288–1290, 2004.

[95] Y. Lu, Y. Huang, and T. Li. More is Different: Constructing the Most Comprehensive Human Protein High-Order Interaction Dataset. *BioRxiv*, pages 2023–11, 2023.

[96] G. Mahalakshmi, S. Ramalingam, and A. Manikandan. An Energy Efficient Data Fault Prediction Based Clustering and Routing Protocol using Hybrid ASSO with MERNN in Wireless Sensor Network. *Telecommunication Systems*, 86(1):61–82, 2024.

[97] S. Maletić and M. Rajković. Combinatorial Laplacian and Entropy of Simplicial Complexes Associated with Complex Networks. *The European Physical Journal Special Topics*, 212(1):77–97, 2012.

[98] N. Malod-Dognin and N. Pržulj. Functional Geometry of Protein Interactomes. *Bioinformatics*, 35(19):3727–3734, 2019.

[99] H. B. Mann and D. R. Whitney. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18:50–60, 1947.

[100] A. Marsden. Eigenvalues of the Laplacian and their Relationship to the Connectedness of a Graph. *Summer Apprentice Paper, University of Chicago, REU*, 2013.

[101] C. T. Martínez-Martínez, J. Méndez-Bermúdez, F. A. Rodrigues, and E. Estrada. Nonuniform Random Graphs on the Plane: A Scaling Study. *Physical Review E*, 105(3):034304, 2022.

[102] X. Miao, K. Liu, Y. He, D. Papadias, Q. Ma, and Y. Liu. Agnostic Diagnosis: Discovering Silent Failures in Wireless Sensor Networks. *IEEE Transactions on Wireless Communications*, 12(12):6067–6075, 2013.

[103] S. Miththapala, J. Seidensticker, L. Phillips, S. Fernando, and J. Smallwood. Identification of Individual Leopards (Panthera pardus kotiya) using Spot Pattern Variation. *Journal of Zoology*, 218(4):527–536, 1989.

[104] T. J. Moore, R. J. Drost, P. Basu, R. Ramanathan, and A. Swami. Analyzing Collaboration Networks Using Simplicial Complexes: A Case Study. In *2012 Proceedings IEEE INFOCOM Workshops*, pages 238–243. IEEE, 2012.

[105] M. Motz, I. Kober, C. Girardot, E. Loeser, U. Bauer, M. Albers, G. Moeckel, E. Minch, H. Voss, C. Kilger, et al. Elucidation of an Archaeal Replication Protein Network to Generate Enhanced PCR Enzymes. *Journal of Biological Chemistry*, 277(18):16179–16188, 2002.

[106] M. N. Mowla, N. Mowla, A. S. Shah, K. M. Rabie, and T. Shongwe. Internet of Things and Wireless Sensor Networks for Smart Agriculture Applications: A Survey. *IEEE Access*, 11:145813–145852, 2023.

[107] A. Muhammad and M. Egerstedt. Control Using Higher Order Laplacians in Network Topologies. In *Proc. of 17th International Symposium on Mathematical Theory of Networks and Systems*, pages 1024–1038, 2006.

[108] A. Muhammad and A. Jadbabaie. Decentralized Computation of Homology Groups in Networks by Gossip. In *American Control Conference, 2007. ACC'07*, pages 3438–3443. IEEE, 2007.

[109] T. Muhammed and R. A. Shaikh. An Analysis of Fault Detection Strategies in Wireless Sensor Networks. *Journal of Network and Computer Applications*, 78:267–287, 2017.

[110] S. Mukherjee and J. Steenbergen. Random Walks on Simplicial Complexes and Harmonics. *Random Structures & Algorithms*, 49(2):379–405, 2016.

[111] M. T. Nair and A. Singh. *Linear Algebra*. Springer, 2018.

[112] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[113] P. Noirot and M.-F. Noirot-Gros. Protein Interaction Networks in Bacteria. *Current Opinion in Microbiology*, 7(5):505–512, 2004.

[114] S. L. Ooi, X. Pan, B. D. Peyser, P. Ye, P. B. Meluh, D. S. Yuan, R. A. Irizarry, J. S. Bader, F. A. Spencer, and J. D. Boeke. Global Synthetic-Lethality Analysis and Yeast Functional Profiling. *Trends in Genetics*, 22(1):56–63, 2006.

[115] S. K. Osterrieder, C. Salgado Kent, C. J. Anderson, I. M. Parnum, and R. W. Robinson. Whisker Spot Patterns: a Noninvasive Method of Individual Identification of Australian Sea Lions (Neophoca cinerea). *Journal of Mammalogy*, 96(5):988–997, 2015.

[116] P. Pamfilos. *Lectures on Euclidean Geometry - Volume 2: Circle Measurement, Transformations, Space Geometry, Conics*, volume 2. Springer Nature, 2024.

[117] A. Panicker and V. Sasidevan. Social Adaptive Behavior and Oscillatory Prevalence in an Epidemic Model on Evolving Random Geometric Graphs. *Chaos, Solitons & Fractals*, 178:114407, 2024.

[118] K. Pearson. VII. Mathematical Contributions to the Theory of Evolution.—III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, (187):253–318, 1896.

[119] M. Penrose. *Random Geometric Graphs*, volume 5. Oxford University Press, Oxford, 2003.

[120] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino. Homological Scaffolds of Brain Functional Networks. *Journal of The Royal Society Interface*, 11(101):20140873, 2014.

[121] J. Polastre, R. Szewczyk, A. Mainwaring, D. Culler, and J. Anderson. Analysis of Wireless Sensor Networks for Habitat Monitoring. In C. Raghavendra, K. M. Sivalingam, and T. Znati, editors, *Wireless Sensor Networks*, chapter 18, pages 399–423. Springer, 2004.

[122] S. Porta, E. Strano, V. Iacoviello, R. Messora, V. Latora, A. Cardillo, F. Wang, and S. Scellato. Street Centrality and Densities of Retail and Services in Bologna, Italy. *Environment and Planning B: Planning and Design*, 36(3):450–465, 2009.

[123] J.-C. Rain, L. Selig, H. De Reuse, V. Battaglia, et al. The Protein–Protein Interaction Map of Helicobacter pylori. *Nature*, 409(6817):211, 2001.

[124] U. Raj and S. Bhattacharya. Some Generalized Centralities in Higher-Order Networks Represented by Simplicial Complexes. *Journal of Complex Networks*, 11(5):cnad032, 2023.

[125] K. Raman. Construction and Analysis of Protein–Protein Interaction Networks. *Automated Experimentation*, 2:1–11, 2010.

[126] S. Ramazani, J. Kanno, R. R. Selmic, and M. R. Brust. Topological and Combinatorial Coverage Hole Detection in Coordinate-Free Wireless Sensor Networks. *International Journal of Sensor Networks*, 21(1):40–52, 2016.

[127] M. V. Ramesh. Design, Development, and Deployment of a Wireless Sensor Network for Detection of Landslides. *Ad Hoc Networks*, 13:2–18, 2014.

[128] R. H. Ramos, C. d. O. L. Ferreira, and A. Simao. Human Protein–Protein Interaction Networks: A Topological Comparison Review. *Heliyon*, 2024.

[129] M. Robinson, C. Capraro, C. Joslyn, E. Purvine, B. Praggastis, S. Ranshous, and A. Sathanur. Local Homology of Abstract Simplicial Complexes. *arXiv preprint arXiv:1805.11547*, 2018.

[130] Y. Rochat. Closeness Centrality Extended to Unconnected Graphs: The Harmonic Centrality Index. In C. Hirschi, editor, *6th Applications of Social Network Analysis Conference (ASNA 2009)*, number EPFL-CONF-200525, Amsterdam, 2009. Elsevier.

[131] S. Roman. *Fundamentals of Group Theory: An Advanced Approach*. Springer Science & Business Media, 2011.

[132] T. Rout, A. Mohapatra, M. Kar, and D. K. Muduly. Essential Proteins in Cancer Networks: A Graph-Based Perspective using Dijkstra's Algorithm. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 13(1):42, 2024.

[133] J.-F. Rual, K. Venkatesan, H. Tong, T. Hirozane-Kishikawa, et al. Towards a Proteome-Scale Map of the Human Protein–Protein Interaction Network. *Nature*, 437(7062):1173, 2005.

[134] M. Sadeghi Ghahroudi, A. Shahrabi, S. M. Ghoreyshi, and F. A. Alfouzan. Distributed Node Deployment Algorithms in Mobile Wireless Sensor Networks: Survey and Challenges. *ACM Transactions on Sensor Networks*, 19(4):1–26, 2023.

[135] U. Saeed, S. U. Jan, Y.-D. Lee, and I. Koo. Fault Diagnosis Based on Extremely Randomized Trees in Wireless Sensor Networks. *Reliability Engineering & System Safety*, 205:107284, 2021.

[136] D. Saha, S. Mitra, and A. Sensharma. Critically Spanning Epidemic Outbreak Cluster in Random Geometric Networks. *Physica A: Statistical Mechanics and its Applications*, 629:129226, 2023.

[137] M. R. Sanatkar, W. N. White, B. Natarajan, C. M. Scoglio, and K. A. Garrett. Epidemic Threshold of an SIS model in Dynamic Switching Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(3):345–355, 2015.

[138] F. A. Santos, P. K. Tewarie, P. Baudot, A. Luchicchi, D. Barros de Souza, G. Girier, A. P. Milan, T. Broeders, E. G. Centeno, R. Cofre, et al. Emergence

of High-Order Functional Hubs in the Human Brain. *BioRxiv*, pages 2023–02, 2023.

[139] M. T. Schaub, A. R. Benson, P. Horn, G. Lippner, and A. Jadbabaie. Random Walks on Simplicial Complexes and the Normalized Hodge 1-Laplacian. *SIAM Review*, 62(2):353–391, 2020.

[140] D. Schoch. *A Positional Approach for Network Centrality*. PhD thesis, Universität Konstanz, 2015.

[141] M. Seringhaus, A. Paccanaro, A. Borneman, M. Snyder, and M. Gerstein. Predicting Essential Genes in Fungal Genomes. *Genome Research*, 16(9):1126–1135, 2006.

[142] D. H. Serrano and D. S. Gómez. Centrality Measures in Simplicial Complexes: Applications of Topological Data Analysis to Network Science. *Applied Mathematics and Computation*, 382:125331, 2020.

[143] D. H. Serrano, J. Hernández-Serrano, and D. S. Gómez. Simplicial Degree in Complex Networks. Applications of Topological Data Analysis to Network Science. *Chaos, Solitons & Fractals*, 137:109839, 2020.

[144] M. Sheerin. *Random Rectangular Networks, Theory and Applications*. PhD thesis, University of Strathclyde, 2018.

[145] M. D. Shirley and S. P. Rushton. The Impacts of Network Topology on Disease Spread. *Ecological Complexity*, 2(3):287–299, 2005.

[146] G. Simionato and M. G. Cimino. Swarm Intelligence for Hole Detection and Healing in Wireless Sensor Networks. *Computer Networks*, 250:110538, 2024.

[147] A. Sizemore, C. Giusti, and D. S. Bassett. Classification of Weighted Networks Through Mesoscale Homological Features. *Journal of Complex Networks*, 5:245–273, 2016.

[148] E. Song. Persistent Homology Analysis of Type 2 Diabetes Genome-Wide Association Studies in Protein–Protein Interaction Networks. *Frontiers in Genetics*, 14:1270185, 2023.

[149] C. Spearman. The Proof and Measurement of Association Between Two Things. *International Journal of Epidemiology*, 39(5):1137–1150, 2010.

[150] M. P. Stumpf and P. J. Ingram. Probability Models for Degree Distributions of Protein Interaction Networks. *EPL (Europhysics Letters)*, 71(1):152–158, 2005.

[151] H. L. Sweeney and D. W. Hammers. Muscle Contraction. *Cold Spring Harbor Perspectives in Biology*, 10(2):a023200, 2018.

[152] M. R. Symonds and A. Moussalli. A Brief Guide to Model Selection, Multimodel Inference and Model Averaging in Behavioural Ecology Using Akaike's Information Criterion. *Behavioral Ecology and Sociobiology*, 65(1):13–21, 2011.

[153] A. Tahbaz-Salehi and A. Jadbabaie. Distributed Coverage Verification in Sensor Networks Without Location Information. *IEEE Transactions on Automatic Control*, 55(8):1837–1849, 2010.

[154] A. Tausz, M. Vejdemo-Johansson, and H. Adams. JavaPlex: A Research Software Package for Persistent (Co)Homology. In H. Hong and C. Yap, editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136, 2014. Software available at `http://appliedtopology.github.io/javaplex/`.

[155] N. Temene, C. Sergiou, C. Georgiou, and V. Vassiliou. A Survey on Mobility in Wireless Sensor Networks. *Ad Hoc Networks*, 125:102726, 2022.

[156] X. Tian, M. Zhu, L. Li, and C. Wu. Identifying Protein–Protein Interaction in Drosophila Adult Heads by Tandem Affinity Purification (TAP). *JoVE (Journal of Visualized Experiments)*, (82):e50968, 2013.

[157] P. Tirandazi, A. Rahiminasab, and M. J. Ebadi. An Efficient Coverage and Connectivity Algorithm Based on Mobile Robots for Wireless Sensor Networks. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–23, 2023.

[158] D. Tocchi, C. Sys, A. Papola, F. Tinessa, F. Simonelli, and V. Marzano. Hypergraph-Based Centrality Metrics for Maritime Container Service Networks: A Worldwide Application. *Journal of Transport Geography*, 98:103225, 2022.

[159] L. Torres, A. S. Blevins, D. Bassett, and T. Eliassi-Rad. The Why, How, and When of Representations for Complex Systems. *SIAM Review*, 63(3):435–485, 2021.

[160] C.-C. Tseng and S.-L. Lee. Identification of Station Importance of Taipei Metro Network using Subgraph Centrality. In *2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 1–2. IEEE, 2021.

[161] F. Tudisco and D. J. Higham. Node and Edge Nonlinear Eigenvector Centrality for Hypergraphs. *Communications Physics*, 4(1):201, 2021.

[162] K. Turnbull, S. Lunagómez, C. Nemeth, and E. Airoldi. Latent Space Modeling of Hypergraph Data. *Journal of the American Statistical Association*, 119(548):2634–2646, 2024.

[163] P. Uetz, Y.-A. Dong, C. Zeretzke, C. Atzler, A. Baiker, B. Berger, S. V. Rajagopala, M. Roupelieva, D. Rose, E. Fossum, et al. Herpesviral Protein Networks and their Interaction with the Human Proteome. *Science*, 311(5758):239–242, 2006.

[164] A. Ullah, F. S. Khan, Z. Mohy-Ud-Din, N. Hassany, J. Z. Gul, M. Khan, W. Y. Kim, Y. C. Park, and M. M. Rehman. A Hybrid Approach for Energy Consumption and Improvement in Sensor Network Lifespan in Wireless Sensor Networks. *Sensors*, 24(5):1353, 2024.

[165] W. Van Criekinge and R. Beyaert. Yeast Two-Hybrid: State of the Art. *Biological Procedures Online*, 2:1–38, 1999.

[166] P. Van Mieghem, J. Omic, and R. Kooij. Virus Spread in Networks. *IEEE/ACM Transactions On Networking*, 17(1):1–14, 2008.

[167] L. Vietoris. Über den Höheren Zusammenhang Kompakter Räume und eine Klasse Von Yusammenhangstreuen Abbildungen. *Mathematische Annalen*, 97(1):454–472, 1927.

[168] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative Assessment of Large-Scale Data Sets of Protein–Protein Interactions. *Nature*, 417(6887):399–403, 2002.

[169] J. Wang, M. Li, H. Wang, and Y. Pan. Identification of Essential Proteins Based on Edge Clustering Coefficient. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1070–1080, 2011.

[170] M. Wei, S. Brandhorst, M. Shelehchi, H. Mirzaei, C. W. Cheng, J. Budniak, S. Groshen, W. J. Mack, E. Guen, S. Di Biase, et al. Fasting-Mimicking Diet and Markers/Risk Factors for Aging, Diabetes, Cancer, and Cardiovascular Disease. *Science Translational Medicine*, 9(377):eaai8700, 2017.

[171] C. Ye, Q. Wu, S. Chen, X. Zhang, W. Xu, Y. Wu, Y. Zhang, and Y. Yue. ECDEP: Identifying Essential Proteins Based on Evolutionary Community Discovery and Subcellular Localization. *BMC Genomics*, 25(1):117, 2024.

[172] M. Younis, I. F. Senturk, K. Akkaya, S. Lee, and F. Senel. Topology Management Techniques for Tolerating Node Failures in Wireless Sensor Networks: A Survey. *Computer Networks*, 58:254–283, 2014.

[173] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLOS Computational Biology*, 3(4):e59, 2007.

[174] H. Zhuge and J. Zhang. Topological Centrality and its E-Science Applications. *Journal of the American Society for Information Science and Technology*, 61(9):1824–1841, 2010.