Identifying Graph Characteristics in Growing Vascular Networks

Christopher Finn Plummer
Identifying Graph Characteristics in Growing Vascular Networks

Christopher Finn Plummer

Abstract

One of the ways that a vascular network grows is through the process of angiogenesis, whereby a new blood vessel forms as a branch from an existing vessel towards an area which is stimulating vascular growth. Due to the demands for nutrients and waste transport, growing tumour cells will access the surrounding vascular network by inducing angiogenesis. Once the tumour is connected with the vascular system it can grow further and colonize distant organs. Given the critical nature of this step in tumour development, there is a demand for mathematical and computational models to provide an understanding of the process for treatment in predictive medicine. These models allow us to generate vascular networks that demonstrate similar behaviour to that of the observed networks; however, there is a lack of quantifiable measures of similarity between generated networks, or, of a generated and real network. Furthermore, there is not an established way to determine which measures hold the most relevance to distinguishing similarity. To construct such a measure we transform our generated vascular networks into an abstract graph representation which allows exploration of the plethora of graph centralities. We propose to determine the relevance of a centrality by finding one that acts as a synthetic likelihood function for estimating the model's parameters with minimal error. Evaluating the relevance of many centralities, it is then possible to suggest which centralities should be used to quantitatively determine similarity. This allows for a way to measure how realistic a model's growth is, and if given sufficient data, to distinguish between regular and tumour-induced angiogenesis and use it within cancer screening.
## Contents

1 Introduction ................................................. 3  
   1.1 Thesis Structure ....................................... 5  

2 Background ................................................... 6  
   2.1 Angiogenesis ............................................ 6  
      2.1.1 Composition of Capillaries ....................... 6  
      2.1.2 Secretion of Angiogenic Factors .................. 6  
      2.1.3 Initiation of Capillary Sprout ................... 7  
      2.1.4 Proliferation and Migration of Endothelial Cells 7  
      2.1.5 Tumour-Induced Angiogenesis ..................... 8  
   2.2 Vascular Network to Graph Representation .............. 8  
   2.3 Centralities and Characteristics ....................... 9  
      2.3.1 Small-World Properties ......................... 10  
      2.3.2 The Four-Dimension Approach of Constructing Centralities 11  
      2.3.3 Reachability .................................... 12  
      2.3.4 Amount of Flow .................................. 12  
      2.3.5 Vitality ......................................... 13  
      2.3.6 Feedback ........................................ 13  
   2.4 Random Network Models .................................. 14  
      2.4.1 Erdős-Rényi Random Graph Model .................. 15  
      2.4.2 Watts-Strogatz Small-World Model ............... 17  
      2.4.3 Barabási-Albert Preferential Attachment Model .... 18  
      2.4.4 Implications to the Methodology .................. 18  

3 Related Work ................................................. 20  
   3.1 Quantification of Sampled Vascular Networks .......... 20  
   3.2 Computational Modelling of Angiogenesis ............... 21  
   3.3 Remarks .............................................. 22  

4 Growth-Branch-Join Model .................................. 23  
   4.1 General Parameters .................................... 23  
   4.2 Growth Parameters .................................... 23  
   4.3 Branching Parameters .................................. 24  
   4.4 Joining Parameters .................................... 24  

5 Methodology .................................................. 25  
   5.1 Constructing the Synthetic Likelihood Function ....... 25  
      5.1.1 Overview of the Synthetic Likelihood Method .... 25  
      5.1.2 Application To Centralities and Characteristic Functions 26  
      5.1.3 Choosing of Characteristic Function ............. 27  
      5.1.4 Evaluation of Parameter Estimation ............... 28  
   5.2 Screening of Centralities and Characteristics ......... 29  

6 Results ....................................................... 31  
   6.1 Analysis of Small-World Properties ................... 31  
      6.1.1 Characteristic Path Length ....................... 32  
      6.1.2 Local Clustering Coefficient ..................... 35  
   6.2 The Four-Dimension Approach for Potential Centralities 37  
      6.2.1 Reachability ..................................... 37  
      6.2.2 Amount of Flow .................................. 37  
      6.2.3 Vitality ......................................... 38  
      6.2.4 Feedback ........................................ 38  
      6.2.5 Summary ......................................... 40  
   6.3 Characteristic Function Implications .................... 40  
   6.4 Incorporating Directional Measures .................... 41  
   6.5 Assumption of Sample Sizes ............................ 41  

7 Discussion .................................................... 43  
   7.1 Limitations ............................................ 44  

8 Conclusion ................................................... 46
1 Introduction

The process of angiogenesis allows blood vessel growth from the branching of an existing vessel and acts as a way for the vascular network to provide nutrients to a cellular region [1]. When a cellular region does not have access to its required nourishment, it can secrete pro-angiogenic factors to its surroundings and signal to nearby existing blood vasculature to construct a new vessel towards the source of malnourished cells [2]. Since cancerous tumour cells grow at a quicker rate and in a malignant manner, they will consume the required nutrients in the surrounding area and the growth will be inhibited. Just as healthy cells secrete a pro-angiogenic factor, tumour cells replicate this behaviour with tumour angiogenesis factors, to induce the creation of a newly formed blood vessel towards it [3]. At this stage the tumour poses a larger threat to the host as it can now grow at an expedited rate without bounds on nutrients and will have the ability to colonize distant organs in the greater vascular network [4]. It follows that with the ability to determine if the behaviour or structure of a vascular network is the consequence of tumour-induced angiogenesis then it is possible to identify the growth at an earlier stage and would allow for a predictive treatment.

Due to this potential and the threat that cancer poses, it has been of relevance to develop mathematical and computational models of the angiogenesis process to allow for their use within predictive medicine [2]. Models generate a sample vascular network through their simulation of vascular growth, however from surveyed papers, time-consuming manual visual inspection remains a primary method to evaluate if the model’s output has the desired behaviour. Hence, it is desirable to find a quantitative method to determine the similarity of sample networks with one another. Or further, to use vascular quantification from real samples to compare with a model’s sample network and deem if they provide realistic behaviour [5]. As such, we want to transform our problem into the domain of graph theory in which we have many metrics that allow us to characterize and respectively compare two graphs [6].

To map from the vasculature to a graph, we can take the generated samples from models and convert them into an abstract graph representation by denoting all the branch points of the vessels as a vertex and an edge between two nodes if the branch points are connected by a vessel in the blood vasculature. This will allow for two interpretations, a directed and an undirected graph which does and does not keep the flow of blood encoded into the network.

As described, we have extracted a directed or undirected graph, $G = (V, E)$, that denotes the branching of our initial blood network, now we can investigate the structure of the network with various centralities. A centrality is a measure on each node of a graph that aims to distinguish how important the node is to the network [6]. Dependent on the measure, it will describe the behaviour or role of a node in the network [8]. An example centrality is the degree of a node, which denotes how many neighbours that the node has. If a node has a high
degree, relative to the other nodes, then it may provide specific functionality and significance in the network, for example a hub in a transport network. Therefore, if we consider two nodes in separate graphs that have a similar centrality value, then we might be able to deduce that these nodes provide the same functionality or that their neighbourhood structure is similar. Consider if we extend this to the general distribution of the centralities, then we can start to deduce comparable functionality or structure throughout both networks and that the graphs are similar. The characteristic function is a function on the set of all centrality values in a graph and provides a description of the distribution of centrality values. And so, we can quantify how similar the networks are by computing the difference of their characteristic functions, described further in Section 2.3. Extending our example, depicted in Figure 2 it is illustrated how a characteristic function is able to provide a measure of similarity between two networks.

![Graphs with different mean degrees](image)

Figure 2: Visually we can see the similarity in structure between the last two graphs and the contrast between the initial graph. Having a characteristic function measuring the mean degree allows for a quantitative measure of this difference.

Now having a way to quantify similarity between the graphs, we want to determine the ‘best’ centrality and characteristic function pair in the context of distinguishing the structure of vascular networks. Ideally we could simply investigate potential centralities and select a single basic centrality, such as the degree, and it would be capable of describing if two vascular networks were similar. However, the complexities of network behaviour can not be reflected by a single measure. Instead, we will require the use of multiple centralities that can measure different aspects of the network and combine them together in our interpretation of the network structure. This raises the question: is it optimal centrality to simply consider as many centralities as feasible? Allowing to measure the similarity of the graphs in as many feasible dimensions. While this may be the case, inevitably we would consider the measure of some centralities to be more important than others. This promises potential to distill the set of possible centralities to the most relevant or ‘basis set’ of centralities.

Given that we want to find the most relevant characteristic function and basis set of centralities, it raises an issue into how to determine the relevance of each centrality to our application. While we can use biological and other domain specific reasoning to propose candidate centralities, we do not have a way to measure how relevant each centrality is. Within Section 5 we introduce the notion of constructing a synthetic likelihood function for each parameter of a computational model. The likelihood function will use the centrality and characteristic function to provide a description of the network for its inference. We can then use the likelihood function to provide an estimate of the model’s parameters. If the centrality is not sensitive to changes in the network from deviation in the parameter, the likelihood function will also not be sensitive and will produce substantial error in its estimation. Therefore, we can use the error of the estimates as a measure of relevance for the underlying centrality. From which we are able to compare various centralities and present a rationale for those that best describe the structure of a vascular network.

Keeping this in mind for a moment, a graph exhibits small-world properties if the neighbours of a node are generally also the neighbours of one-another and if the average distance from each node to another is small. These can be measured with the global clustering coefficient and characteristic path length centralities, respectively. Small-world properties and the corresponding centralities are further described in Section 2.3.1. Notably, many biological networks exhibit small-world behaviours [9]. As highlighted in Section 3 we are motivated by the dependent use of the small-world properties in the evaluation of model behaviour. The reasoning provided is that if a computational model creates networks that demonstrate small-world properties, then it is some measure of correctness. However, there is no further reasoning as to why this is the best measure for a vascular network and there exists a need into the comparison of other quantification measures of vascular networks [5].

As such, we will analyze the respective small-world centralities with the provided methodology and use the Growth-Branch-Join (GBJ) model for parameter estimation. Section 4
describes the GBJ model further, at a high-level it aims to provide a simplified model of angiogenesis by dividing it into three key stages: growth, branching and joining of simulated vessels. By simplicity, the number of model parameters is reduced which allows for the conceptual separation between them as the correlation between parameters is less convoluted. Further, the scale of the model generates sample networks that are large enough to provide interest, and also, the simplicity allows us to reduce the computational cost of generating each sample. This will prove important due to the number of samples required in the construction of the synthetic likelihood functions.

Having described our methodology, we then evaluate the small-world properties with it to establish a benchmark. Using these benchmarks we further explore the evaluation of other potential centralities such that we are able to contrast them with one another. During this exploration a number of concerns are raised that we subsequently address so as to further establish the applicability of the methodology. Having done so, we then aim to show whether the small-world properties are capable and should be recommended to be used as a quantitative measure of vascular network similarity. Or, if there are other centralities that are more favourable to be used as such.

1.1 Thesis Structure

Initially in Section 2 we establish the required background knowledge in the various domains of the thesis. Namely, a biological description of angiogenesis, image analysis of vascular networks, as well as, an overview of network centralities and random network models. Section 3 provides a survey of recent efforts in computational modelling and quantification of vascular networks. This provides primary motivation and reasoning into the study of the small-world properties and the consideration of other biological motivated measures. Outlined in Section 4 is the Growth-Branch-Join model, the model is used in the evaluation of the following methodology and discussion. Section 5 presents the methodology of constructing synthetic likelihood functions and described the desired properties of centralities. Then, Section 6 enumerates the results that were obtained and establishes a platform for the proceeding discussion. Section 7. Finally, the conclusion summarizes our findings and provides thoughts on future implications.
2 Background

2.1 Angiogenesis

The process of angiogenesis encapsulates how a new blood vessel is created from an existing vessel and can be broken down into several stages. Enumerated as: the secretion of pro-angiogenic factors, initiation of a capillary sprout, the proliferation and migration of endothelial cells (ECs) and further maturation of the new vessel. In this section we will provide a short description of capillary composition and the aforementioned steps to provide understanding for biological motivation and future discussion.

2.1.1 Composition of Capillaries

Functionally, capillaries are primarily required to provide a smooth surface for blood flow and permeability to allow exchange of nutrients and waste between the blood and the surrounding tissues. To provide a smooth surface there is a layer of ECs in a tube shape, called the endothelium, which lines the blood vessel. Enveloping the endothelium layer is a thin structural layer that provides support for the shaping of the capillary within the extracellular matrix (ECM) called the basement membrane (terminologically also referred to as the basal lamina). Given the thin nature of the endothelium and basement membrane, it allows for the desired permeability. The last main component is the pericytes embedded around the ECs in the basement membrane. Pericytes provide multiple functions for the capillary: additional support in the vessel lining, control over permeability and maintenance of the chemical balance of the ECs. We can visually this structure as depicted in Figure 3.

![Figure 3: Visualization of capillary composition.](image)

2.1.2 Secretion of Angiogenic Factors

As described, capillaries allow for nutrients and waste to permeate through the cell wall with the surrounding tissue. However, tissue must be within a close proximity to the capillary to receive nourishment. Surrounding regions that do not have sufficient nutrients require that the capillary network is extended into their region, this is the primary functionality of angiogenesis. To induce angiogenesis from the surrounding vasculature, a deficient region will secrete pro-angiogenic factors (AFs). We can imaging this diffusion as the overlay of the cells in Figure 4. As mentioned, pericytes control the chemical balance of the ECs, one of which is a chemical balance between pro- and anti-angiogenic factors. When the balance is in favour of the pro-angiogenic factor, it will initiate the angiogenic process on the ECs. Further, the deficient tissue or cellular region secretes the pro-angiogenic factor and it diffuses outwards creating a gradient of factor between the vasculature and region.
2.1.3 Initiation of Capillary Sprout

Once the pro-angiogenic factor has initiated angiogenesis at a particular location of the blood vessel several processes begin. The structure of the ECs changes to increase permeability and the vessel dilates. The consequential increase of nutrients and outward pressure facilitates the ability for the production of proteolytic enzymes which will degrade the surrounding basement membrane \[1\]. The ECs located where the membrane is degraded will become activated into tip endothelial cells (TECs) and will create a gradient of anti-angiogenic factor laterally along the vessel to prevent other TECs from forming nearby \[2\]. Another property of TECs is their development of filopodia, which are “slender cytoplasmic protrusions” (Vilanova et al. \[4\]) that acts as receptors for the diffused factor and remnants of the basement membrane. The other non-tip ECs in the region will also become activated as stalk endothelial cells (SECs) and are key to the proliferation of ECs in the following stage. The TEC will act as this initial sprout of the capillary into the ECM.

Figure 5: Depiction of the initial sprout of the TEC as it goes into the ECM through the area of degraded membrane. Additionally, we see the lateral anti-angiogenic factor gradient to prevent other tip cells from forming in the proximity and the filopodia that form at the endpoint of the TEC.

2.1.4 Proliferation and Migration of Endothelial Cells

For the sprout to propagate through the ECM, it is required that a path is cleaved through it. This is done by the proteolytic enzymes that also initially degraded the basement membrane. After which, the TECs are able to migrate through the ECM. As previously mentioned, there is the gradient of secreted AF and the filopodia are able to navigate the TEC along the gradient towards the initial source. SECs will proliferate and recruit ECs into the space behind the TEC, facilitating the elongation of the sprout \[2\].

Figure 6: Here we have the proliferation of the endothelial cells in the form of stalk endothelial cells that fill the space behind the TEC. We can also see the pericytes will recruit and follow along the growth sprout.

This describes the basic mechanism for the growth of the capillary sprout, however, there are notable behaviours that we can further describe as they are commonly considered in mod-
els. As mentioned, filopodia are also capable of detecting broken down basement membranes in the region. Areas with remains of a basement membrane are considered to have less resistance to migrate within, hence the filopodia will direct the TEC towards them [4]. This results in a TEC to intersect with another TEC or the sprout of a TEC and results in the sprouts merging, called anastomosis. Creating loops in the developing network is crucial to potentially create blood flow [3]. The other behaviour that leads to loops is further branching of the capillary sprouts. As described by the lateral inhibition of angiogenic factor, a similar gradient occurs along the length of the new sprout. As the sprout elongates the decrease in anti-angiogenic factor will causes the same process to occur on the proliferated ECs [1]. This results in branched sprouts that will also follow the same process of migration towards the source of AF.

2.1.5 Tumour-Induced Angiogenesis

As we have described, the use of pro- and anti-angiogenic factors are crucial to the initiation and activation of a new blood vessel forming. When we consider how a tumour can induce angiogenesis it will produce similar pro-angiogenic factors called tumour angiogenic factors (TAFs). It can also be the case that the tumour will induce surrounding cells in the ECM to become malignant and also secrete additional factors [1, 2]. In a similar process as described before, angiogenesis will occur and the capillary sprouts will migrate towards the tumour’s region.

2.2 Vascular Network to Graph Representation

Vascular networks are created through complex cellular processes, and as a result, detecting if two networks provide a similar structure or functionality is a difficult task. To do so through visual inspection is nearly impossible, instead we rely on transforming the domain from an image into another representation. The use of image analysis techniques is an approach that helps to automate parts of this process, however as will be mentioned, there are key steps that prove to be extremely difficult to automate and hinder its capabilities. We are primarily concerned when transforming the representation to that of an abstracted graph, \( G = (V, E) \), however in the study of vascular network quantification, measures can also be taken directly from the image analysis, for example the total area of the image that is taken up by vessels [5, 10]. Let us give a brief overview of how the features of a graph can be used to represent the vascular network and a high-level outline of how we could ideally convert an image into a graph representation, as well as, the difficulties that prevent this.

At the core of the representation is what constitutes the vertices and edges, in our context. Branch points of vessels and the existence of a vessel connecting two branch points denote vertices and edges, respectively. Generally there is not a use of vertex weights, however, the edges can have many different interpretations [3, 11, 12]:

- **Vessel Length**: The edges can be extended to include weights that model the vessel length.
- **Blood Flow Direction**: By using a directed graph, the edges can be extended to model the direction of blood flow throughout the network.
- **Blood Flow Volume**: By extending the edges to include a weight that denotes the diameter of the vessel, it is possible to model the volume of blood flow through the network.
- **Arc-Chord Ratio**: By extending the edges to have weights denoting the arc-chord ratio, it is possible to model the tortuosity of the network. The arc-chord ratio is the ratio of the length of the vessel over the distance between the two branch points, and is a measure of tortuosity of a curve, where a tortuous curve exhibits many turns or bends within it.

Using one or a combination of these allow the representation to keep some of its topological features embedded into the abstracted graph. Having these different interpretations also allows for the same centrality to measure variable aspects of the underlying network. These also provide motivation into the types of centralities that can be considered or constructed.

While the extraction of a graph from a vascular network will need to be specialized to the vast types of samples being considered, it will generally be a variation on the following image analysis pipeline. The primary goal of extraction is to be able to differentiate between the vessels and the background. First, there will be pre-processing techniques to help create
a consistent lighting environment. Then there will be the binarization of the image using a threshold algorithm. This results in a new binary image where ideally each pixel with value 1 represents a vessel and 0 represents a background value. Commonly there can be some dilation of the binary images to remove blemishes from possible image artifacts. Then the core of the pipeline is to create a skeleton of the binarized image, also referred to as a thinning algorithm. A skeleton is a binary image such that is reduced to a single-pixel wide representation, which ideally models the topological structure of the network. This step would have to then be extended to try and approximate any of the aforementioned edge weights. From here we can determine the branch points and edges between them. We should note that there is increased difficulty when we want to measure any of the mentioned topological features. This difficulty is a result of the vasculature imaging having vastly different compositions, and so when considering the length, diameter or tortuosity of a vessel, how to measure it is not clear and requires manual specialization. However, the without these measures there is a loss of information about the network. This is why it is not possible to have a general fully automated image analysis pipeline, but we can still use components of it to help when we transform a network.

Figure 7: Expanded markings from Figure [1] to also include the edges (blue) and endpoints (red) with the branch points (green).

2.3 Centralities and Characteristics

Network science consists of the broad study of how networks behave and are structured, including what has been previously described, how to convert objects from the biological (or physical) world into a network or graph representation. Once we have the network there
are many ways that we can measure or interpret the behaviour, one of the primary ways of measurement is the notion of a centrality. A centrality seeks to provide a quantification of how important each node of a network is; depending on what the centrality measures, it will define what is important and can highlight a different structure or functionality of a network. This is important to the interpretation of the measures, and so, as we introduce the following centralities, we will illustrate what their measurement indicates as important. When we compute the centralities of a graph, we can also consider the general behaviour of all the centralities, which is generally described as a characteristic or property of the network. For instance, we might want to check if all of the nodes behave similarly for a centrality, if there are nodes that exhibit extremes, or, the distribution of the centrality values. When we can measure the behaviours of a network in terms of a characteristic and the underlying centralities, we now have a method to discern how similar two networks are with each other. Further, we can define a difference of their respective behaviour in a quantitative manner.

In a formal sense, we will denote a network as a set of nodes (vertices) and edges, \( G = (V, E) \), and then each centrality, \( C \), works as a function on each of the nodes, \( u \in V \), in the network. This allows for a way to quantify the importance of the node with respect to the property imposed by \( C \),

\[
C : V \rightarrow \mathbb{R}
\]

We will denote a characteristic function of a graph to be a function on the set of all its centralities,

\[
F_C(G) : \{C(u) | \forall u \in V\} \rightarrow \mathbb{R}^n,
\]

for example the mean value. Note that this will differ with other definitions of a network characteristic, however, it will prove useful to narrow the scope done to a function that has a range of \( \mathbb{R}^n \). In particular, then we can construct a concrete definition of the difference between characteristics. This will act as the basis of the similarity between two networks, \( G \) and \( H \), with respect to a centrality, \( C \):

\[
|F_C(G) - F_C(H)|
\]

where a smaller difference of Euclidean distance implies increased similarity. We can extend this further to consider the normalized difference or any other measure of difference as our characteristic is now a function mapping to \( \mathbb{R}^n \). We can also introduce the notion that a characteristic function can be a combination of various other characteristic functions. If we have a set of characteristic functions, \( \{F_i\} \) such that each has their respective dimensionality of \( d_i \) then we can let a characteristic function simply be all the appended values into a single output vector with a dimension \( d = \sum_i d_i \). For instance, we would have one characteristic function be the mean of all the centrality values and another be the standard deviation. Then both of characteristics are of the form described in Equation 1 with \( n = 1 \) and we could append them such that we create a new characteristic with \( n = 2 \) such that the first component is the mean and the second is the standard deviation. Allowing us to encompass more behaviour into a single characteristic value.

Having defined a description of a centrality and characteristic functions, as well as, how we can use these to compare the likeness of graphs, we have the tools required to construct our methodology. However, we have not provided explanations for different types of centralities and what they measure. In the following sub-sections a description of all referenced centralities in the thesis are described. We focus on providing a description along with the implications or meaning that a measure would imply in the context of a vascular network. Due to the importance of the small-world properties we first highlight these measures, then we discuss the other centralities grouped by similar functionality or importance measure.

2.3.1 Small-World Properties

The small-world properties encompass the behaviour of three different centralities of the graph, namely the degree distribution, the characteristic path length (CPL) and the global clustering coefficient. As will be discussed further in Section 2.4 and discovered in Section 3, these properties appear prominently in real networks and there is evidence that vascular networks also exhibit these properties. Further there is a dependence on these properties for the evaluation of computational models. Hence, we put a primary focus on these properties at this point. Before we discuss the type of behaviour let us define these centralities [6] [8].

- The degree of a node, \( u \), measures how many edges have the node as an endpoint, denoted as \( k(u) \). When we consider a directed graph, we denote a difference of when the edge begins at \( u \) or ends at \( u \). Then then we can differentiate between the out-degree and in-degree by counting the number of out- or in-edges respectively. This is
the most basic measure of importance of a graph, as it is effectively counts the number of neighbours that the node has. In this sense, a node with a high-degree would be perceived as more important as it is connected to many other nodes, or it can ‘reach’ many other nodes easier. In our networks, this denotes the number of vessels that are merging or branching at the branch point.

- The characteristic path length (CPL) of a node denotes the average shortest path length from the given node to all other nodes in the network. When a graph is not connected, and \( u \) and \( v \) are in different components, then it is common to let the distance be \( \infty \), however, in our case we will modify the definition such that we do not sum for the node pairs that are in different components. Let us denote \( d_{uv} \) as the shortest path between two nodes of the graph, \( u \) and \( v \). Then the CPL of the node \( u \) is defined as

\[
\langle d_{uv} \rangle,
\]

where \( u, v \) are in the same components of the graph. The definition follows in the same way for a directed graph except that when we compute the shortest path it is not possible to travel in both directions on the edges. Further, when we consider a graph with weighted edges, then we may have the weights of the edge be the distance between two nodes in our path. If we consider the vascular network as a transport network of nutrients, then this allows for a measure of how effectively the node can transport nutrients to other nodes in the network. Of note, this measure does not follow the general guideline that a centrality should have a higher value denote more importance, as in this case a lower CPL would denote more importance as it could reach other nodes more effectively.

- The global clustering coefficient is the ratio:

\[
\frac{3 \times \text{(number of triangles)}}{\text{(number of connected triples)}},
\]

where a connected triple is defined as any set of 3 nodes such that exists at least 2 edges between them. We can also consider the local clustering coefficient of a node, \( u \) is given as the ratio of:

\[
\frac{\text{(number of neighbour pairs that are connected )}}{\text{(number of possible edges between all of } u \text{'s neighbours)}}.
\]

This measure of the local clustering is also extended to take the average of all these values and is also referred to as the global clustering, hence we will be deliberate which measure we are considering when we use it as a centrality. These clustering centralities are measures of the transitivity in the network, where we can say a node, \( u \), is transitive by having the property that if another node, \( v \), is a neighbour of \( u \) and \( w \) is a neighbour of \( u \) then \( v \) is a neighbour of \( w \). Given this transitive nature, they were largely motivated in the study of social networks, and the interpretation that if I know someone, then there is a higher likelihood that I know the people they know. Given the spatial dimension of our vascular networks, if one of the branch points, \( v \), is connected to another branch point, \( u \), and then \( u \) is connected to another branch point, \( w \), then it would be more likely that \( v \) and \( w \) are also connected due to their physical proximity.

As we discussed, these centralities allow us to determine the importance of the nodes with respect to one another in a graph, but we can apply characteristic functions to characterize behaviours of the network. The small-world properties are primarily a combination of the network having a low CPL and a high global clustering coefficient. It can also be extended, and commonly is, to also require the network to be a scale-free network. Where a scale-free network is characterized by having the degree distribution following the shape of the power law distribution familiar from statistics \[6\]. In the context of network science, the distribution will be dominated by many nodes that have a small degree and then a long-tail of increasing degrees. With this in mind, we have that the combination of these attributes allows for the formation of hubs of highly connected nodes that are connected with each other and the large portion of the low degree nodes are connected to these hubs directly.

### 2.3.2 The Four-Dimension Approach of Constructing Centralities

One of the approaches to the consideration of potential centralities is based on the 4-dimension approach proposed by D. Koschützki et al \[13\]. The 4-dimension approach describes a method.
for the construction of a new centrality that is specialized to a given application. An important aspect of this approach is the use of a basic term, of which, the core idea is that we can, in general, categorize a centrality into one of the following: reachability, amount of flow, vitality and feedback. As such, it is important to consider measures from all of these categories when we are exploring a new application. With this in mind, the following sub-sections describe the properties of each category and a description of our investigated centralities.

2.3.3 Reachability

We can consider a centrality to measure reachability if it is a measure of how central the node is within the network, or in other words, its ability to ‘reach’ other vertices. This generally contains degree or distance measures and we consider the CPL, efficiency, degree, eigenvector and Katz centralities. As such they are defined as [6, 8]:

- The efficiency, also referred to as the closeness, is a variation of the CPL measure. It is the reciprocal value of the CPL, and so the efficiency of a node \( u \) is

\[
\frac{1}{\langle d_{uv} \rangle}.
\]

Generally, we prefer a centrality value to have a higher magnitude as an indication of importance, when we considered the CPL this was not the case as the nodes with a minimal CPL had the most reach. Furthermore, when a graph is not connected it is the case that the CPL is defined as \( \infty \) or you are limited to computing the distance for nodes in the same components, if we then want consider a characteristic function with any summation of the node values we would have that it would be \( \infty \). In the efficiency, these then becomes a value of 0 and so the centrality is more robustness to such networks without requiring a loss of information, and so it is better suited for such characteristic functions, therefore we would generally prefer this measure. Physically it has the same interpretation of the CPL.

- The eigenvector centrality builds on the degree centrality by considering its greater neighbourhood proportional to the network. We let \( A \) denote the adjacency matrix of \( G \), then we can compute the eigenvector of \( A \), \( \lambda \). Then the eigenvector centrality is defined for node, \( u \), as the \( u \)-th element of \( x \) such that \( x \) solves

\[
Ax = \lambda x.
\]

This arises from the intuition to make a node important if it has a high degree or if it is connected to neighbours that have high degrees. Recalling the interpretation of the degree to be the number of vessels that are merging or branching from the branch point, then this is extended to be the proportion of the the vessels in the network that branch from the node or its neighbours.

- The Katz centrality also builds on the degree centrality to consider the greater neighbourhood, instead of computing the entire proportion, instead we provide an attenuation factor, \( \alpha \), and a basic degree value \( \beta \). Then we further extend the eigenvector centrality by solving for \( x \) in:

\[
x = \alpha Ax + \beta
\]

In this way we provide another measure that lets a node be important if it has a high degree or if it is connected in a close proximity to nodes that are have a high degree.

2.3.4 Amount of Flow

The amount of flow measures how much of the network flow goes through a given node. When we consider the vascular network as a transport network, we would consider this as what proportion of the nutrient transportation flows through the node at one point. We can measure this by assuming that nutrients would be transported along the shortest path to its destination, and so we can measure how often a node is in the shortest path from node to node. The shortest path stress and betweenness centralities are measures in this category that we consider:

- The stress centrality measures the total number of shortest paths that a node, \( u \), is apart of. This is computed by enumerating all the possible shortest paths from each pair of nodes, \( s \) and \( t \). Then we compute the number of these paths that contain \( u \), where \( u \neq s \neq t \) is not the start or end point. It is important to note that there may
be many shortest paths between a node pair, and that all of these are included in the enumeration. As we have described, this physically allows for a measure of how much of the transportation of nutrients in the network would flow through the node.

- The betweenness centrality is the intuitive extension of the stress centrality to normalize it with respect to the number of other shortest paths that exist between $s$ and $t$. As such, the centrality for a node, $u$, would first enumerate the paths as in the stress centrality, then for each of the shortest paths between $s, t \neq u$ such that $u$ is apart of, divide by the total number of shortest paths between $s$ and $t$ and sum these fractional values. Hence giving more importance to the shortest paths that are more vital in the sense that if node were removed from it then there would be a noticeably increase in the transport difference. It has the same physical implications, however, we know have an increased importance for nodes that are apart of more vital shortest paths.

2.3.5 Vitality

A vertex, $u$, is considered to be vital with respect to $F_C$ if we see a relatively big difference in the measure when we remove $u$. Formally, we would denote the centrality value for a node, $u$, as

$$F_C(G) - F_C(G \setminus u).$$

This can be generally applied with any characteristic function, but we focus on the application of the radius and diameter of a graph, defined below. Behaviourally this would allow for measures of how important a node is to the connectivity of a network. However, we need to compute the characteristic function of the entire network and then recompute this value for each node of the network to determine the difference, since the sample networks range from 100-1 000 nodes, it is unfortunately deemed as computational infeasible for our method.

Instead we can consider the similar measures of percolation and robustness of a network. These use similar methods of removing nodes in a network and studying the effect on the network when removing them, percolation does so randomly where as a robustness will select a node of significance. Percolation would require similar computational requirements as we would want to study the removal of many random nodes. However, we can provide a centrality of robustness that will simply remove the node with the highest centrality within the network and measure the difference in the centralities of the graph without such a node. With this in mind, we consider the robustness of the network with respect to the radius and diameter measures.

The radius and diameter are different measures of the eccentricity of a network, these are defined as:

- The eccentricity of a graph is another measure based on the shortest paths of a graph, for a node, $u$, it measures the maximum of the shortest path distances between $u$ to any other node, $v$, in the network:

$$\max\left(\{d_{uv} \mid \forall v \in V\}\right).$$

In this case we do want to let $d_{uv} = \infty$ if they are in different components as it will allow us to detect if the removal of a node caused the network to become disconnected. In this way, we can use it in combination with vitality, percolation and robustness to measure the connectivity of a network, or in other words, determine how redundant nodes are.

- The radius is then a global measure that returns the minimum value of the eccentricity and the diameter returns the maximum value of the eccentricity. As noted, these can then be used as a measure of vitality or redundancy of a node, respectively.

2.3.6 Feedback

A feedback centrality is one that uses an underlying centrality, $C$, and their values. Then for each node we can compute a measure on the set of centralities for each node. This differs from a characteristic function as it can also provide this measure per node rather than globally. Given the rather wide description, many of the centralities that provide a global measure of a graph can fall into this category. For instance, we can consider the cluster coefficients to be within the feedback category as it is measure based on the number of triangles in the neighbourhoods of each node.

We are interested in measures that use the degree distribution as the underlying centralities, as this appears to be a strong characteristic of vascular networks. In real vascular
networks all branch and join points would only consist of a vessel branching into two or two vessels merging into one. Hence, the structure of the extracted network would have branch points of degree 3, or end points of degree 1. In this manner we are motivated to have measures that are measures of how close to regular a graph is, where an $n$-regular graph is one such that each node has a degree of $n$. In this manner, we introduce the measure of assortativity, regularity and entropy as considerations in the feedback category. They are defined as:

- The **assortativity coefficient** is quantified by the probability that a node, $u$, with a high degree has neighbours that also have a high average degree. Which is to say that an associative graph will have its high-degree nodes connected with other high-degree nodes. In contrast, a dissociative graph will have that nodes with a high degree will connect with other nodes of a low-degree. To construct a measure of this first consider $p_k$ denote the probability that a random node of the network has degree $k$, further, consider instead of choosing a random vertex, we choose a random edge and then use the vertex following the edge. Since edges have a higher proportion to the nodes of high degree, as they have an increased number of edges, then the probability grows proportional to $k p_k$. So we define $q_k$ to denote the probability that if we choose an random edge and then select the node at the end of the edge, it will have a remaining degree of $k$, where the remaining degree is the number of edges on the node excluding the chosen edge, so the node would have a degree of $k + 1$. Then we can define the distribution of $q_k$ as [14]:

$$q_k = \frac{(k + 1)p_{k+1}}{\sum_j j p_k}.$$  

Furthermore, we want to denote the joint probability that from the chosen edge, they have respective degrees, $j$ and $k$, denoted as $e_{jk}$. We can then combine this into a measure of assortativity as if $e_{jk} = q_j q_k$ then the network is assortative, or disassortative, and we will require a normalizing value to allow for comparison across networks. This is given by the variance, $\sigma_q^2$ and we can define the assortativity coefficient as [14]:

$$\frac{1}{\sigma_q^2} \sum_{jk} j k (e_{jk} - q_j q_k).$$

- The **regularity** of a graph, is a modified version of the degree centrality to incorporate a measure of regularity by normalizing the distance of a node to the average degree of the network. In this manner, if a node, $u$, satisfies, $k(u) = r$ then it has a centrality value of 1 and quickly depreciates. We define the following centrality:

$$-\frac{k(u)^2 - 2r k(u)}{r^2},$$

where $r$ denotes the average degree of the network. As has been noted, we then provide a measure of how far the node is to the regular value of the network, and if we take the average of all nodes we get a measure of how regular the graph is as a whole.

- The **entropy** of a graph aims to provide a measure of uncertainty of a node, in the sense of how predictable the neighbourhood of the node is. In the context of a regular network such as a lattice, we would note that all nodes would have a low entropy as each of the neighbourhoods would be identical. Therefore, we would expect the hierarchical or lattice-like vascular network to have a low entropy. To compute a global measure we first compute the degree centralities of the network, then we can construct a probability that a random node in the network has a degree of $k$, $p(k)$. Using the formula for Shannon’s entropy we can compute it as:

$$- \sum_k p(k) \log p(k),$$

for all the present degrees, $k$, in the network.

### 2.4 Random Network Models

Having described how we can characterize and measure the behaviours of a network, it allows for models that generate networks exhibiting these given behaviours. As with any model, it aims to provide a simpler description of the real behaviour, which allows for an easier
understanding of the dynamics in the growth and distribution of the network. Further, the study of simpler network models has been motivated due to the difficulty in providing mathematical descriptions and consequential study of real networks. Hence, it would prove necessary to construct simple models that could emulate the behaviour of real world networks. Of which, it would be possible to define formal mathematical descriptions and allow for the study of their mathematical properties. Initially, we can consider the pioneering work of Rapoport, Gilbert, Erdős and Rényi to establish the foundational concepts of a random graph model. Here we will demonstrate an example of mathematical analysis of networks and how this is used to demonstrate a small characteristic path length. Building on this model, we will introduce two alternative approaches to create networks with small-world properties that are not present in the initial model of Erdős and Rényi. Namely, the initial random graph model does not generate local clustering behaviour and the degree distribution does not resemble that of a scale-free network. The following section then provides background insight into the foundational models of the research field and how the models progressed to emulate real-world networks through the small-world properties. It also acts as context to the motivation into the relevance of the model’s parameters which will be used within the methodology.

2.4.1 Erdős-Rényi Random Graph Model

The initial efforts to construct models of networks were made independently through the work of Rapoport, Gilbert, Erdős and Rényi [9, 15]. Where they describe similar methods to construct a network, \( G \), through the use of randomly selecting the edges between nodes of the graph. Given the prevalence of the Erdős-Rényi model, we will use their description in which we can denote the model to generate a graph, \( G_{n,p} \), such that it consists of \( n \) nodes and that there is an edge between any pair of nodes \( u, v \) with probability \( p \) [9]. As we have mentioned, this provides a much simpler graph to reason and understand the dynamics of. For instance, it is then possible to provide rigorous descriptions of the probability of each node to have degree \( k \) and the expected mean degree of all nodes in the network which would be intractable from networks generated through a biological phenomena. Respectively these can be denoted and described as:

\[
p_k = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{and,} \quad \bar{d} = (n-1)p,
\]

where we use \( n - 1 \) because we don’t allow for self-loops in the network [9] [15]. It is also then of interest to investigate how the generated networks behaviour changes as the model parameters are deviated. With reference to the empirical study of Avrachenkov and Dreveton we will first describe the behaviours of the random graph model as \( p \) is deviated and then we discuss the implications [15]. As such, we will let \( \alpha \) be some constant and then for a given \( n = 100 \), let \( p_n = \frac{\alpha}{n} \). Hence, the average degree is

\[
d_{n,\alpha} = (n-1)p_n \approx np_n = n\frac{\alpha}{n} = \alpha
\]

where for sufficiently large \( n \) the approximation is negligible. It was then investigated how the network would form components which is quantified by measuring the ratio of nodes that were in the largest component. Through empirical data, it was then shown that for smaller values of \( \alpha < 1 \) the network consisted of many isolated nodes or very small components. In contrast, when \( \alpha > 1 \) the proportion of the graph that is in the largest component increases dramatically and converges to all nodes being within the largest component. In the same manner, if we let \( p_n = \alpha \frac{\log n}{n} \) and then we have

\[
d_{n,\alpha} = (n-1)p_n \approx np_n = n\alpha \frac{\log n}{n} = \alpha \log n.
\]

It is then further demonstrated that for \( \alpha < 1 \) the probability that the entire network is close to 0. However, immediately when \( \alpha \to 1 \) and further increases the probability that the network is connected converges to 1. Finally, in the data we had denoted that \( n = 100 \) but we can note that for a larger \( n \) the same behaviours will occur however the respective increases after \( \alpha > 1 \) will occur at a dampened rate. These results are also similarly enumerated by Newman wherein the only difference is the mean component size and largest component sizes are the metrics used [9].
(a) Fraction of the nodes belonging to the largest connected component of the graph with respect to $\alpha$, where $p_n = \frac{\alpha}{n}$ and $d_{n,\alpha} = \alpha$.

(b) Empirical probability that the graph is connected with respect to $\alpha$, where $p_n = \alpha \log n$ and $d_{n,\alpha} = \alpha \log n$.

Figure 8: We can see empirically that we have a phase transition for $\alpha = 1$ of the Erdős-Rényi random graph model. As was demonstrated in Figure 2.3 of the work of Avrachenkov and Dreveton [15] (CC BY-NC).

With this in mind, let us consider the implication of this behaviour. Intuitively, being able to show that the measure of a model is sensitive to changes in the parameter then denotes that the parameter is relevant to the corresponding network structure that is measured. In this case, the increased probability of edges to connect correlates to both: a larger ratio of the nodes being in the largest component and a higher probability that the network is connected. In this example it is clear, however when we later consider the GBJ-model with many parameters it will be more convoluted to determine how the parameters correlate to the network structures with respect to different measures. But, this provides the underlying reasoning. However, we also have a more complex behaviour in that when $\alpha = 1$ there is a dramatic change in both of the values. Before and after this point the networks have a distinctively different behaviour and this is referred to as a phase transition point. We transition from the phase $\alpha < 1$ with many isolated nodes into the phase of $\alpha > 1$ where there is a large giant connected component, hence $\alpha = 1$ is the phase transition point. Finding such behaviour allows for a better understanding and ability to predict the structural properties of the network and is important when developing further models. For instance, when we now want to investigate the CPL of the random graphs we have the information that we need to let $p_n$ be relative to $n$ such that we can ensure our graph is connected and the measure is not ill-defined.

With this in mind, we can consider the computation of the CPL with reference to Newman [9]. Recalling that $d = (n - 1)p$, then if we want to consider the mean number of nodes that are $x$ neighbours away from a node, it can be given as $d^x$. So if we want to approximate the average distance to any other node we would need it such that $d^\ell \simeq n$. Which we can compute as

$$\ell = \frac{\log n}{\log d}$$

and is considered to be small relatively [9]. Based on the behaviour of how the edges are constructed and the independent sampling of each edge, it follows that the neighbours of a node are not any more likely to be connected to the other neighbours then they would be with any other node. Inherently, we would expect this to decrease the amount of clustering in the network and that most nodes would have a similar degree, therefore the network would have a lower mean clustering coefficient and the distribution of degree would not follow a scale-free network. This is shown to be the case, the clustering can be computed as $\frac{\ell}{n-1}$ which vanishes as $n$ becomes large and the distribution of the degrees follows that of a Poisson distribution [6, 9]. Hence the Erdős-Rényi random graph model does not exhibit all the properties of a small-world network.

This has allowed us to demonstrate the capabilities in the analysis of the mathematical or statistical properties of the generated random networks that are feasible due to the simplified model. It shows that we can provide a reasoning of relevance for parameters with respect to the measures and through the existence of a phase transition point in the structural behaviour it allows for demonstration in the ability for the simplified models to retain dynamics. It is also then extended to be able to show that the model is capable to produce graphs that contain a small CPL, however, it did not contain the clustering and scale-free properties that would be required for all the small-world properties.
2.4.2 Watts-Strogatz Small-World Model

The Watts-Strogatz small-world model was created with a motivation to be a simple model that retained a high clustering coefficient and a small CPL. As such, it aims to address the first of the aforementioned short-comings, particularly for networks that can scale to any size. The most straightforward way to create a scalable network with a high clustering coefficient would be with different types of a lattice [6]. However, the CPL of a lattice would be quite large as there are no ‘short-cuts’ to traverse from one side of the network to the other efficiently. However, as has been discussed the random graph did show to have a property of a smaller CPL and the Watts-Strogatz model takes advantage of this to construct a model that, in essence, interpolates between these two types of networks, a lattice and a random-like graph.

We will initially start with a single-dimension circular lattice, depicted in Figure 9, such that there are $n$ nodes and each node is connected to the $c$ nodes next to it [9]. As the name implies, this holds the properties of a lattice network and has a high clustering coefficient but a large CPL. Then we for each edge in the network it will be rewired with a probability $p$, wherein each rewired edge will be removed from the graph and a new edge will be placed between any two randomly selected vertices. This describes the original small-world model, but it is of note that there are some variations of this model that do not remove the edges when rewiring and will just add the new edge to help with analytic calculations [6].

![Figure 9: As described in the original paper of the Watts-Strogatz model [16], we can see from their illustration (Figure 1) how the dynamics of the network change as we increase the amount of rewiring, defined by $p$. Copyright 1998 by Springer Nature. Reprinted with permission.](image)

For thorough reasoning and all the computation of these measures we refer to Chapter 15 of Networks: An Introduction [6], but we can highlight the key results of the model. As we had varied $p$ in the random model, we can investigate the behaviour of the small-world model networks when $p$ is varied. Expectantly for $p = 0$, the lattice has a high-clustering coefficient by also a high CPL and for $p = 1$ the graph is random-like and retains a low-clustering coefficient but a large CPL. What is interesting, is that when we increase $p \to 0.1$ from 0 we see the CPL decrease at a substantial rate due to the introduced ‘short-cuts’ that are added to traverse quickly from one side of the network to the other. Furthermore, since $p < 0.1$ is still close to 0 these new edges will be a very small portion of the mean and so the clustering coefficient does not decrease at the same rate. At $p = 0.1$ we see that the CPL is close to its minimum and converges toward it and so the rate in which it decreases is very small for $p > 0.1$. In contrast, the proportion of edges that are randomly selected now begins to increase and so the clustering coefficient begins to decrease substantially for the same reasons as described in the random graph. Hence, there exists a region when $0 < p < 0.1$ such that the generated networks have a high-clustering coefficient and low CPL, characteristic of a small-world network. However, the other issue mentioned previously is the degree distribution of the network. Clearly when $p = 0$ the network is a lattice with each node degree being equivalent and does not exhibit the scale-free network property. Hence, $p = 1$ the network is random-like and does not exhibit the scale-free network property, similarly when $p = 1$ the network has the same properties to the random graph model which we discerned as not having the scale-free network property. Since for all $0 < p < 1$ it is an interpolation between these two types of networks, it also follows that they do not have a degree distribution associated with a scale-free network.

As was perhaps foreshadowed by initially stating the goals of the small-world model were, we unfortunately still do not have a model that exhibits all of the small-world properties due to the lack of being a scale-free network. However, we have described a model that provides a simple method capable of generating models with a high-clustering coefficient and low CPL.
2.4.3 Barabási–Albert Preferential Attachment Model

Having been provided models that are capable to emulate the first two small-world properties, we then want to consider if there is a mechanism to construct networks such that they are scale-free. In this regard, we consider the proposed Barabási–Albert model, commonly refereed to as the preferential attachment model, due to its simplicity and the intuitive nature in which it emulates the required power-law distribution of a scale-free network. As the name implies, the model looks to create a network such that nodes will prefer to attach to particular nodes over others, arising from the behaviour in real networks [17]. Namely, in real networks it appears that there are hubs that are very well connected and that when growing, new nodes prefer to connect to the hubs that are the most connected. In contrast to the previous models described which created fixed number of nodes at the initial step, the Barabási–Albert model aims to model the growth of the network by initializing an arbitrary starting graph, $G_0$, followed by adding a single node to the network at a time. Then when the node is added it will connect probabilistically to $m$ other nodes based on how connected they are. Where $m < m_0$ and $m_0$ is the number of nodes in $G_0$. To do so the node will connect to another node with probability proportional to its degree. Letting $k_u$ denote the degree of node $u$ then the probability that the new node would connect with node $u$ is

$$\Pi(k_u) = \frac{k_u}{\sum_{v \in G} k_v}$$

which is repeated $m$ times. There are some variations here about how to do this sampling process, in particular if we want to allow for self-loops or multiple edges between nodes. For simplicity of the model it is accepted to allow for both of these with the reasoning that as the number of nodes tends to infinity, then these edges become negligible [17].

![Figure 10: A sample network generated by the Barabási–Albert random network. The model graphs are characterized by the formation of hubs as we have described in a scale-free network and this is also shown quantitatively in the distribution of degrees throughout the network.](image)

With this in mind, we can consider the behaviour of the model networks with respect to the small-world properties. The diameter of the network will grow slower than $\ln N$ which is slower than that of the random graph from Equation 2 [17]. Hence, it does have low CPL as desired. While the clustering coefficient provides more local clustering than that of a random graph, it still does not have the corresponding high clustering coefficient associated with a real small-world network [17]. Most importantly however, when we consider the distribution of the nodes throughout the network it is shown that they following a power-law distribution of order $\gamma = 3$. From the nature of the algorithm we have that as we grow the network all of the newly connected nodes would have less time and be much less likely to have any other nodes connect to them, hence we see a rich-get-rich phenomenon which is indicative of a power-law distribution. Therefore, it is possible to create a model with a simple mechanism to construct networks that display a power-law distribution in their degrees and denote a scale-free network.

2.4.4 Implications to the Methodology

In the preceding sections we have motivated and illustrated the construction of three prolific network models. Due to their simple description it helps to provide understanding of what behaviours can emerge from the models and that exist in real networks. In the same vein, it has also shown how we can couple the parameters of a network model and the sensitivity of centralities to such parameters. For example in Figure 8, we considered how the constant, $\alpha$, would correspond to the ratio of how many of the nodes were in the largest component. Let us define this ratio as a characteristic function, $F_C \in \mathbb{R}$ and recalling $p_n = \frac{\alpha}{n}$, then we can think of this as a mapping from $\alpha$ to a generated graph, $G_{n,p_n}$, and further to the value of its characteristic value $F_C(G_{n,p_n})$. Naturally, given the bijective form of the mapping in Figure 8.
we could then think to consider the reverse mapping. Providing a way to estimate $\alpha$ given the characteristic value $F_C(G_{n,p_n})$, it would not be an exact answer due to the probabilistic nature of the model. Further if we consider how accurate the estimate is, it would be dependent on the slope of the bijective mapping, or in other words, the sensitivity to perturbations in the parameter that is being estimated at a given point. Hence, we can conclude that if $F_C(G_{n,p_n})$ is sensitive to changes in the parameter $\alpha$ then it is a relevant measure of the behaviour that the parameter is modelling. Likewise, we could employ a similar reasoning to denote that the CPL and global clustering coefficient provide a characteristic function that is relevant to the measure of the parameters of the small-world model. For the sake of contrast, if we consider a characteristic function that computed the number of nodes in the network, it would not be able to estimate any parameter other than $n$ in either model due to the static nature of $n$. In this extreme, we see that it is not sensitive to the underlying changes of $\alpha$ and hence it is not relevant to this measure. All this to say, that we are able to provide a metric for the relevance that a centrality and characteristic function have for a given model and its parameters. While these follow nicely in these simplified models, such relations are not apparent in more complex models with increased parameter dimensionality. However, the underlying reasoning provided from the description of these models is the same that will be used when we are in the domain of computational models of angiogenesis. This reasoning will be further elaborated in the Methodology and we can reference here for the ‘base case’ example.
3 Related Work

Having established the required knowledge and an initial motivation to the methodology in Section 2, we have the required knowledge such that we can describe the current work and its implications in the related area. Additionally, when constructing the aim of the thesis it is important to first survey the current state of research in related fields to ensure that we are not providing redundant results. Furthermore, surveying will allow us to provide motivation as to why we should do our research and insight into the direction to consider what aspects are important in the field.

In the context of network modelling of vascular growth, we can split the related work into two domains: the quantification of \textit{in vivo} and/or \textit{in vitro} vascular network samples and constructing mathematical models of angiogenesis. Work towards the quantification of vascular structure allows us to determine which measures have already been considered and have prior biological motivation either for or against their use. Within the study of computational models is the discussion of methodology to evaluate potential models. Such discussions aim to highlight what they interpret as the key behaviours or structures of a vascular network and are used when displaying that their respective model emulates them. Further, these behaviours are used as a qualitative metric of evaluation which we can use as insight for the evaluation of a computational model. As such, in the following section we provide an overview of current efforts in both domains and how we can combine the work from both to help us formulate the desired insights of this thesis.

3.1 Quantification of Sampled Vascular Networks

When considering the aim in the current research in the quantification of (micro)-vascular networks, we see a consistent theme to discover measures that are able to distinguish between networks of a different nature. For the sake of example we can consider the metric that computes the total vessel area in an image. As will be further described below, it has been shown that we can use such a measure to help discern between a network where tumour-induced angiogenesis has occurred and a control sample network in which it has not \cite{10}. Having such a measure allows for the automation in comparison of networks, as well as, highlights a common use-case and need for such metrics. As a result, many possible metrics have been generated, however, only a subset are considered correctly normalized and standardized \cite{5}. In this manner we mean a metric to be normalized if it can be applicable to networks that occur at different scales and standardized if it is acknowledged as applicable across the research community. With reference to our example, we would want to not only consider the total vessel area of the image but also ensure that we make the measure relative to the scale of the image. Providing a proper normalization of a metric is critical to its ability to compare differing models that may occur at different scales. The evaluation of when a metric becomes standardized appears to be rather opaque and is more a test of time than a given method to evaluate a metric quantitatively \cite{5}. Being able to provide such a method would be able to aid in this process. Coupled with this research is the associated development of software to automatically derive such metrics from images and reduce the manual labour required. The possibility that advancements in these automation processes would allow for the use of many alternative methods shapes the requirement for a level of abstraction over the different metrics in our methodology, which is addressed through the mapping of metrics to centralities and characteristic functions. With this in mind, we will provide a survey of various papers to illustrate how and which metrics are being investigated.

As we alluded to earlier, we return to our example of total vessel area. The study investigated at the structure of the vascular network that was created due to angiogenesis induced by a Glioblastoma tumour \cite{10}. Glioblastoma is a cancer that originates from a group of cells in the brain or spinal cancer, in this case, the study implanted a Glioblastoma tumour into the brain of a mouse. After a period of 3-14 days, images of the tumour micro-vasculature were taken, similarly, images from the same brain region of a control sample of mice were also taken. Of which, they were able to automate the computation of the following morphometric features: total vessel area, number of junctions, vessel length and number of endpoints. Described as:

- Total Vessel Area: Area of the image pixels that the vessels cover.
- Number of Junctions: The total number of branching or merging of the blood vessels. Corresponds to the number of nodes in a graph.
- Total Vessel Length: Euclidean distance over the pixels between the two junctions or endpoints
- Total Number of Endpoints: Number of vessels that did not join to any other vessel yet.

Given that these measures were relative to the pixels, and that the imaging allowed for precise distance of each pixel, the applicable measures were also normalized. These measures were shown to all be capable of distinguishing between the sample and control mice. The growth of vessels in the tumour-induced network created the described behaviour of angiogenesis of vessels branching, merging and continued new vessel growth, which each of the mentioned metrics are direct measures.

As a method to compare the vascularization mechanisms of vascular mimicry (VM) with angiogenesis, A. Fouladzadeh et al. provided an outline for analysis of a vascular network with network-based approaches. A study was constructed with an assay of endothelial colony-forming cells (ECFCs), and two variants of VM assays. The growth of the assays was recorded over a 48-hour time period and images were taken at sample increments; these samples were then transformed to a network representation. Motivated by the general sentiment that many biological networks exhibit properties of a small-world network, the study investigated this property of the assays. As a result, there was evidence that over time, the ECFC network would maintain its property of being a small-world network as the CPL decreased and the global clustering coefficient increased over time. In contrast, the VM networks would have their global clustering coefficient decrease at a quick rate and so they would not maintain a small-world network. Additionally, when incorporating the measure of edge length and thickness it was noted that all the sample networks behaved similarly to traffic networks, notably with respect to the Braess paradox. Briefly, the Braess paradox denotes that if we introduce a new edge between vertices that the average flow throughout the graph can be reduced. Having the study show that both the CPL and global clustering coefficient was able to distinguish networks formed through ECFCs and the VM networks a motivation to the consideration of these measures in our analysis. Additionally, given that it was also shown that the networks resemble networks with similar networks to that of traffic networks indicates that we should consider measures of flow as having particular relevance.

There have been other efforts to use more unique morphometric features such as lacunarity or toposity in network analysis. Lacunarity being a measure of how well patterns or fractals are able to fill the space and toposity as was defined in Section 2.2 In our case, the lacunarity would be measured as a ratio of the pixels that are covered by both the image of a vascular network and an overlayed fractal. Namely, we can consider the work encompassed in studying vascular networks exhibit space-filling fractal behaviour [18, 19]. However, much of the success has been demonstrated at higher levels of vascular structure rather than at the capillary level scale of which tumour-induced angiogenesis would occur. However, at the capillary level it is still apparent to see space-filling behaviour and it motivates that we can construct or consider possible network properties that take into account these other measures. Regardless, research into these unique metrics illustrates the breadth in domains that we can consider measures from. In fact, there are many other potential measures that have not been reiterated due to the quantity, but reference can be found from the mentioned survey work [5].

3.2 Computational Modelling of Angiogenesis

Within mathematical modelling of angiogenesis, the primary difference of models can be broken down to a targeted cellular scale of the model [2]. The cellular scales are enumerated as: sub-cellular, cellular and tissue scale; the computational complexity of a model for a given cellular scale is imperative to the choice of model. Models can be categorized as being continuous, discrete or hybrid and all have had success in describing an aspect of angiogenesis, [1, 2]. For instance, when the study of angiogenesis began with mathematical models, only tip-cell migration was considered and it wasn’t until later that blood flow modelling was also incorporated [20]. The future of research lies in modelling additional aspects of angiogenesis, in particular with the advent of hybrid models that are now able to model at multiple cellular scales. With this in mind, we want to consider how such models are being evaluated. Overall, this process is has been largely done through qualitative measures and descriptions of behaviour and the primary quantitative metric is whether or not the network exhibits small-world properties. So to avoid being repetitive we will consider the following models but they are representative of the models listed in both the survey resource [1, 2].

Inspired by previous study of fractal modelling within tumour vasculature, Chimal-Eguña et al. aim to provide an improved discrete fractal model of vasculature construction [21]. As part of the paper, it was desired to analyze the networks generated by the algorithm by comparing chosen properties with real sampled networks. To do so, the authors were motivated by the investigation of real world networks by Newman [9], and used small-world
properties to determine if the network structure was similar. Namely, the networks were measured for the CPL and global clustering coefficient in a similar manner as previously described, as well as, checking if degree distribution exhibits a scale-free network. Relying solely on these three properties for their evaluation.

An instance of using the phase-field approach to mathematically model angiogenesis is provided by G. Vilanova et al. in [4]. It is a mixed model that uses continuous partial differential equations to model the TAF distribution and discrete agents to model the sprouts. To evaluate the behaviour of the model, the paper provides tuned simulations to emulate a chosen in vivo experiment. Foremost, an evaluation of filopodia sensitivity on network behaviour is presented. Networks generated from varied filopodia sensitivity are compared and the authors use measures such as the number of bifurcations, anastomoses and TECs; these can be described in terms of our network representation as the number of nodes with an out-degree greater than 1, in-degree greater than 1 and out-degree of 0, respectively. In the context of their model, the difference of these measures represent the behaviour of the model having all TECs growing in parallel and not intersecting. Further, the number of loops and vessel lengths are also used to quantify how the networks differ. These measures are analogous to a small-world characteristic and further demonstrate vascular networks exhibiting small-world properties. However, when the authors further demonstrate their model with another experiment they are unable to provide a quantitative comparison and the evaluation of the model is only provided through visual inspection.

3.3 Remarks

As a result of surveying related efforts and research, there appears to be a large dependence on the idea that biological networks exhibit small-world properties. Furthermore, the primary motivation for the use of these properties is simply that biological networks exhibit it. This demonstrates the need for better and more concrete reasoning as to why these properties are more relevant, or to provide evidence that there exists measures that are ‘better’ to evaluate vascular network growth similarity. The latter is further motivated by the suggestion that only basic metrics have been applied to the graph abstraction of micro-vascular networks, and that many techniques exists and have not been explored [5]. These allow us to explore other types of networks metrics, traffic flow, as mentioned, new centralities, redundancy, or information diffusion to name a few [5].
4 Growth-Branch-Join Model

This section is concerned with introducing the Growth-Branch-Join (GBJ) model that is later used as a point of implementation and investigation of the thesis’ main proposed methods. In order to demonstrate the capabilities of the synthetic likelihood method, we are required to implement the method for a giving model of which the parameters we can estimate. While we could potentially choose any computational model that has been provided in the literature, we use the GBJ model due to its stated aim of providing a simplified model of angiogenesis, as well as, the computational requirements of the method used. Given it does not model any specific component of angiogenesis we are able to investigate more general properties of the vascular networks as the parameters are less specific. Furthermore, the model is capable to generate the required number of samples that are used in the method within a computationally feasible manner, another result of its simplicity. Finally, the model has the potential to be extended further if a particular aspect of the vascular network would be of interest. These reasons make the GBJ a good candidate for the use of evaluation inside of the synthetic likelihood method.

To achieve reduced complexity, the model will break down the growth of vessels into 3 steps applied iteratively, growth, branch and join. Briefly, the growth step will elongate each of the blood vessels, then each vessel will attempt to branch and finally vessels that are within each other’s proximity will join. The simulation begins by placing numerous active TECs and sources of AF within a unit square. These can be placed anywhere, but in our simulations we will have the TECs spread along the x-axis and the AF placed in two location on the y-axis. Then the aforementioned steps are applied iteratively until there is no more active TECs.

Additionally, the model uses 10 parameters which are generally orthogonal to one another. In this sense, they don’t model the same processes which allows for an easier separation and interpretation of their behaviour in the model. In the following sub-sections we provide a short description of each model parameter.

4.1 General Parameters

To construct the simulation we have some general parameters to construct the space that the model is within.

\textit{diam}: Each cell will have a diameter as denoted by \textit{diam}. The model constructs the graph of vessels in the unit square of cells.

\textit{\sigma_{diam}}: Add variation to the diameter with a sample standard deviation of \textit{\sigma_{diam}}. Allows to model that cells do not have a uniform size.

\textit{N-Dist}: Denotes the Notch-Delta distance which simplifies this to the number of cell diameters away that the TEC needs to be before it is capable to branch again. As was mentioned in Section 2.1.4, we need a mechanism to prevent branching immediately surrounding branch points.

\textit{\lambda}: Vary the decay in which the AF disappears. As the AF is the primary driving force behind the growth of the TEC towards a certain region, it is important to consider how the concentration of the factor will impact growth.

\textit{AF}^{x,y}: Set the location of the source of AF. Provides the capability to set different locations of the AF in the unit square. Given the location, a concentration gradient of angiogenic factor (AF) is constructed solving

\[ \nabla AF = -\lambda AF + S_{[x,y]} \]

4.2 Growth Parameters

To encapsulate the growth of a single sprout we consider:

\textit{v_0}: Speed of vessel growth. Encapsulates the speed in which the proteolytic enzymes degrade the surrounding ECM and the proliferating ECs migrate into the space. The parameter itself is in terms of time steps.

\textit{\sigma_v}: Standard deviation in the directional noise for the growth of a vessel at each time-step. Naturally, the growth will not proceed in a direct path and will have some wavering direction as it grows. Allows for a way to model this noise.
4.3 Branching Parameters

As the forming vessel continues to elongate away from its own branch point and surpasses the Notch-Delta distance, the model allows the possibility to branch again. The branching behaviour is described by:

$P_{\text{branch}}$: The probability for the TECs to branch. Of note, this is sampled after each growth step of the vessel and must be considered when we vary the growth parameter.

$\theta_{\text{branch}}$: Internal angle of the branch when it occurs.

$\sigma_\theta$: Variation in the exact branch angle. The parameter is the standard deviation in each occurrence of branching.

4.4 Joining Parameters

As the TECs are growing towards the AF it may be the case that they will come within a proximity with one another and as a result the TECs of each will combine back into a single growing set TECs. The model takes this into account with a single parameter:

$J_{\text{dist}}$: Denotes how close two of the TEC need to be for them to join together. The proximity is modelling by the number of separating cells.

(a) Visualization of vessel growth
(b) Additional visualization of branching and anastomoses

Figure 11: Example simulation of a vascular network with default parameters
5 Methodology

As has been outlined, we have transformed our problem domain from a vascular network to an abstract graph representation, \( G \), denoting the branching points of the network. Operating on a graph allows access to a substantial number of possible centralities that we might consider, however, we need to identify a method to determine how much relevance to assign each centrality. Suppose it is possible to estimate a model’s parameter from the measured centrality values of a network generated by the model. Then it follows that the network behaviour, with respect to the centrality, is sensitive to the changes in the estimated parameter. Since the parameter models behaviour within the growth of the network, the centrality is assigned relevance to measuring the similarity of such networks.

To address this, we propose constructing a synthetic likelihood function for the parameters of a computational model; the function can then provide an estimate range for the model’s parameters given a sample network. In the original description of the synthetic likelihood method, the time-series is reduced to a set of summary statistics that described the behaviour of the time-series. Instead we propose the use of a proposed centrality and a characteristic function to model the key structure of the network, acting as the summary statistics, and provides a method to measure the capability of the centrality to estimate a model’s parameter. Furthermore, we can then evaluate the estimation with generated test data from the model and this provides a quantifiable measure of the capabilities to estimate the parameter, and hence a measure of relevance for the centrality.

Having a way to evaluate the centralities and characteristics allows us to determine which are the most relevant. However to explore all the possible centralities and characteristics using the proposed method would be computational and time-infeasible. Hence we then also introduce a method of screening measures for proposed centralities to determined if they should be considered.

5.1 Constructing the Synthetic Likelihood Function

We are interested in how a centrality and characteristic function are able to capture the structure of a network, as proposed we want to provide a measure of this based on its ability to estimate a model’s parameters from an observed network. In statistics, a maximum likelihood estimation (MLE) is used to estimate the parameters of a statistical model given some observed data \( \hat{\theta} \). It does so by constructing a likelihood function, \( L(\theta \mid X) \), that denotes the likelihood of the parameter set, \( \theta \), given the observed data, \( X \). Then either analytically or numerically, it is investigated which value \( \hat{\theta} \) provides the maximum value of \( L(\hat{\theta} \mid X) \). The construction of the likelihood function is based on the probability density function (pdf) or probability mass function (pmf) of the parameters, in our context however, the complexity of the models of angiogenesis make it infeasible to get or assume the underlying distribution of these parameters, particularly as the dimensionality increases. This is of course not limited to our context, but to the general use of complex statistical models of processes and has resulted in the research of methods to approximate the likelihood function \( \frac{1}{3} \). In particular, we make use of the synthetic likelihood (SL) method as described by Wood \( \frac{2}{4} \). At a high-level, the SL method looks to approximate the likelihood function by operating on summary statistics of the raw data. Using summary statistics of the underlying data such that they roughly have a multivariate normal distribution allow for construction of a SL function similarly to how we construct a traditional likelihood function. Furthermore, from the SL we are able to estimate the model parameters in the same way as a regular likelihood function. In the following we will first provide an overview of Wood’s SL method and then describe how we can use the method within our context.

5.1.1 Overview of the Synthetic Likelihood Method

Let us describe the construction of the log SL function through the following steps, with close reference to \( \frac{2}{4} \):

- Define the summary statistics: A conversion from the raw input data, \( y \), to a vector of summary statistics, \( s \). It is assumed that we can use a multivariate normality approximation:

\[
s \sim N(\mu_\theta, \Sigma_\theta)
\]

where \( \mu_\theta, \Sigma_\theta \) are the mean vector and covariance matrix of the model parameters, \( \theta \), respectively. Let \( d \) denote the dimensions of \( s \).
• Generate synthetic data: Use the model to create the data sets, \( \{y^*_i\} \), for the desired number of samples \( i = 1, \ldots, N \). Further, convert these into the corresponding summary statistic vectors, \( \{s^*_i\} \).

• Estimate \( \mu_\theta, \Sigma_\theta \): From the generated data, we can compute an estimate of:

\[
\hat{\mu}_\theta = \sum_i \frac{s^*_i}{N},
\]

and,

\[
\hat{\Sigma}_\theta = \frac{SS^T}{N-1}
\]

where \( S = \{s^*_i - \hat{\mu}_\theta\} \) is a \( d \times n \) matrix.

• Construct the Synthetic Log Likelihood Function: Using the assumption of multivariate normal distribution we can then use these to create a log SL function with respect to \( \theta \):

\[
L_s(\theta) = -\frac{1}{2}(s - \hat{\mu}_\theta)^T \hat{\Sigma}_\theta^{-1}(s - \hat{\mu}_\theta) - \frac{1}{2} \log |\hat{\Sigma}_\theta|
\]

Then we have constructed the log SL function for a value of \( \theta \), note that since the log function is monotonic then if \( \theta \) is the maximum \( L_s(\theta) \) it is also the maximum of the underlying likelihood function. It is also important to note that in the given method \( \theta \) can be any dimension and so it could ideally estimate all of the models parameters. However, we start our analysis by investigating each parameter individually and have all other parameters fixed to a default value. As will be further discussed, it would be possible to establish a similar process where we estimate more of the model parameters with a centrality.

This outlines the general structure of using a SL function, so then it follows to describe how we will apply this to our domain. We will outline the process that we follow when constructing the set of SL functions for a candidate centrality and characteristic function: \( C \) and \( F_C \).

5.1.2 Application To Centralities and Characteristic Functions

For reference, Figure 12 describes the architecture of the system and we will provide the detailed steps as follows in the section.

First let us familiarize with the notation that we will use for generating the data samples, \( \{y^*_i\} \), from a model. The model will have a set of parameters (or arguments),

\[ A = \{A_1, A_2, \ldots, A_m\} \]

where \( A_i \) is a closed interval of feasible values for the parameter, as well as, default values \( a_i \) for each parameter.

Now if we consider constructing the SL functions for one of the parameters, \( \theta \), it will have a corresponding closed interval of the feasible parameter values, let \( \theta_{\text{min}} \) and \( \theta_{\text{max}} \) denote the extrema of the interval. Then we can specify a granularity for the parameter interval, \( \alpha_\theta \), to sample over, creating a discretized set of the interval,

\[ \Theta = \{\theta_{\text{min}} + k\alpha_\theta\}, \text{ for } k = 1, \ldots, n \]

where the size is \( n = \frac{\theta_{\text{max}} - \theta_{\text{min}}}{\alpha_\theta} \).

A single sample of the network is denoted as \( G_\theta \) where \( \theta \in \Theta \), and all other parameters are fixed to their default. To generate the samples for \( L_s(\theta) \) we then create an \( N \)-element sample vector with samples from \( \theta \):

\[ \{G_\theta\} \]

The next step is then to compute the centralities of each network

\[ \{y^*_i\} = \{C(G_\theta)\}. \]

Furthermore, from the definition of a characteristic function with respect to a centrality, it is analogous to a summary statistic. So we have that

\[ \{s^*_i\} = \{F_C(G_\theta)\}, \]

where the range of \( F_C \) is \( \mathbb{R}^d \). We will use the term characteristic function for the remainder of the text, however, it is interchangeable with the summary statistics for centrality distributions.
Generate samples over Θ
Compute \( F_C(\{y^*_i\}) \)
Estimate values of \( \hat{\mu}_\theta, \hat{\Sigma}_\theta \)
Compute \( L_s(\theta) \)
Compute \( F_C(y) \)
\( \{y^*_i\} \)
\( \{s^*_i\} \)
\( \hat{\mu}_\theta, \hat{\Sigma}_\theta \)

Figure 12: Visualization of the workflow to compute \( L_s(\theta) \) for a given parameter \( \theta \in \Theta \).

5.1.3 Choosing of Characteristic Function

Using these definitions, we have applied the general SL method to use networks and their centralities as the ‘raw’ data. However, we have not addressed the assumption that we can use a multivariate normality distribution and the possible choices of characteristic functions. Addressing the former first, we can consider that for many types of the characteristic function, it will follow that the central limit theorem (CLT) will apply as \( n \to \infty \) \[24\]. Further as is stated by Wood, “[\( L_s(\theta) \)], is invariant to reparameterization and is robust to the inclusion of uninformative statistics, so very careful selection of [characteristic function] is not necessary” \[24\]. This allows for some required flexibility when we consider the choice of characteristic functions. Otherwise, it is suggested we can apply a transformation to the distribution of the characteristic function to allow for its use. Hence, when we consider the use of a characteristic function it will be necessary to address these concerns.

Further, the distribution of different centrality values over a network can vary drastically. For example, as depicted in Figure 13, the distribution of degrees in a vascular network, in contrast with the distribution of characteristic path lengths. Again iterating that we must consider the underlying distributions when we want to choose a characteristic function.

With that in mind let us motivate the general characteristic function that is used in our evaluation, and when it is applicable to use. We will define our characteristic function as:

\[
F_C(G) = \{\mu, \sigma, \hat{\mu}, c_0, c_1, c_2, c_3\}\]  

where \( \mu \) is the mean, \( \sigma \) is the standard deviation, \( \hat{\mu} \) is the median, and \( c_i \) are the coefficients of the cubic polynomial regression of the sorted centrality values. By the central limit theorem, \( \mu \) will be approximately normally distributed for large \( n \). If the underlying distribution is heavily-tailed, -skewed or multi-modal then we need to be cautious of our value of \( n \). Furthermore, due to our sufficient number of underlying samples, and their general similarity to \( \mu, \sigma \) and \( \hat{\mu} \) will also tend to approximate a normal distribution \[24\]. The polynomial coefficients also generally have an approximate normal distribution, and this can be transformed to only measure on particular quantiles to better fit the normal distribution. Hence, if we do some preliminary analysis of the centrality distributions and check that it does not exhibit any of the mentioned extreme behaviours we can use the multivariate normality distribution. Given that each of the individual measures provide approximate normality, it follows that the combination of them will to.

The choice of these measures within the characteristic function allows us to provide a cohesive quantification of the centrality distribution shape. The general choice a cubic polynomial was made experimentally, however, this is more of a place-holder for any dimensional

(a) Sorted distribution of the degree centrality.

(b) Sorted distribution of the CPL centrality.

Figure 13: Demonstration of different distributions of centralities in a vascular network.
and the estimate, creating the test sample, we are then able to measure the error between the actual parameter and the estimate, $|\hat{\theta}_s - \theta_s|$, of the parameter estimation. If we let $|\Theta|$ denote the length of the range of the parameter, then we introduce the range error percentage which measures the average error over all the test samples as a percentage of the parameter range, allowing for a better comparison over parameters. Defining the range error percentage as:

$$\frac{1}{|\Theta|} \sum_s \frac{|\hat{\theta}_s - \theta_s|}{N_s} \times 100\%. \quad (9)$$

As a point of reference, let us consider a $L_s(\theta)$ to be a function that randomly produces an estimate $\hat{\theta}_s$ by generated a random value in $\Theta$. Then we want to consider what the expected range error percentage would be. Since $\theta_s$ is also a randomly selected value in $\Theta$, we are equivalently computing the expected mean distance between two uniformly random points on a line segment. This is defined as $\frac{d}{3}$, where $d$ is the length of the line segment [25]. So in this case $L_s(\theta)$ as an expected range error percentage of

$$\frac{1}{|\Theta|} \frac{|\Theta|}{3} \times 100\% = 33.3\%. \quad (10)$$

This will allow us for a point of reference when we compare the performance of our SL functions as a threshold for when it is incapable of detecting any changes in the underlying network structure.

Additionally, we are able to plot an SL function, $L_s(\theta)$, over $\Theta$ to get further insight into their behaviour. To help describe such behaviour let us consider what attributes an ideal SL function would have for comparison. To illustrate this consider the following three functions that are examples of the extreme behaviours:

$$L_1(\hat{\theta} | X) = \begin{cases} 1, & -0.25 \leq \theta \leq 0.25 \\ 0, & \text{otherwise} \end{cases}, \quad L_2(\hat{\theta} | X) = \cos(4\theta), \text{ and, } \quad L_3(\hat{\theta} | X) = -\theta^2 + 1,$$

where $\theta \in [-2, 2] = \Theta$ and let the actual parameter be $\theta = 0.05$, depicted in Figure 14. Recalling that our estimate for a parameter is one that corresponds to the maximum value of the SL function, in all three functions the maximum value is given when $L_i(\theta | X) = 1$. For $L_1$ this is true for all $-0.25 \leq \hat{\theta} \leq 0.25$ and so we would not be able to do any better than randomly guess the parameter over that range. This is due to a lack of local robustness to the parameter at the global maximum which would cause for an estimate that is not precise. If we consider $L_2$, we see that around $\theta = 0$ we are able to distinguish that it is a better estimate than its local surroundings. However, we have three points over $\Theta$ that we can not discern between as an optimal estimate, in other words, it has a high modality. Again this would lead to estimates that have significant errors. Finally considering $L_3$, it is immediately evident that its form address both of these issues by begin locally robust at the global maximum and being uni-modal (low modality). With this in mind, we would see that over many samples,
would hold the lowest expected estimate error. Of course, in practice the function will not follow such contrived examples, however this allows us to gain an intuition of the key behaviours that would be desired from our SL functions, namely a high robustness around the global maximum and being uni-modal.

Figure 14: Examples of SL functions that demonstrate the key behaviours that are or are not desired.

In summary, we have provided an overview of constructing a synthetic likelihood function and then further described how we can use the method in our context for parameter estimation of complex stochastic models. We then described how we can evaluate the model and provide a quantification of its relevance with respect to each parameter, \( \theta \). This provides a method for analysis of centralities and their capabilities in parameter estimation.

5.2 Screening of Centralities and Characteristics

In Section 5.1 we provided the methodology for evaluating the capabilities for a centrality, however, there is a discrepancy between the computational time to simulate and compute summary statistics on a network vs time-series sample from the original application and of a generic computational model of angiogenesis. Hence our computational requirements to construct the SL functions will be substantially higher. As a primary goal of the thesis is to determine the centralities that show the most capability at parameter estimation, we want to consider as many centralities as possible. The computational burden to evaluate a single centrality poses a significant challenge. Hence we are motivated to find a way to avoid using the required computation and manual time of inspection to evaluate centralities that would provide promising results, the following screening measures are provided so as to reduce the overhead for the initial consideration of a centrality.

Let us first consider a centrality and characteristic function pair, \( C, F_C \), such that they allow for the following assumptions:

- For any \( \alpha_\theta > 0 \) and \( s > 0 \), such that for all \( k = 1, \ldots, n \) it holds:
  \[
  \bigcup \{ F^s_C(G_{\theta_{min} + k\alpha_\theta}) \cap F^s_C(G_{\theta_{min} + (k+1)\alpha_\theta}) \} = \emptyset,
  \]
  \( (11) \)
  where \( F^s_C(G_{\theta}) \) denotes a set of \( n \) samples.
- For any \( \epsilon > 0 \), there exists an \( \alpha_\theta > 0 \) and \( s > 0 \), such that for all \( k = 1, \ldots, n \) it holds:
  \[
  \phi \left( F^s_C(G_{\theta_{min} + k\alpha_\theta}), F^s_C(G_{\theta_{min} + (k+1)\alpha_\theta}) \right) < \epsilon
  \]
  where \( \phi \) is an operator that takes the minimum distance between any combination of the sets extrema.

Informally we can describe assumption 11 as, for any granularity of \( \Theta \) and sample set size, neighbouring sample sets will not intersect with one another. Likewise, assumption 12 describes the ability for neighbouring sample sets to be arbitrarily close to each with respect to the \( \phi \) operator. Combined, the assumptions allow for a description of continuity over the space of sample sets. When we then consider constructing our set of likelihood functions over \( \Theta \) when we can assume these properties. Namely, if there is no intersection between neighbouring sets then the likelihood functions will have a form similar to a normal distribution with a small standard deviation, and be extremely sensitive around \( \theta \). Having continuity over the range will then also allow for accurate interpolation between likelihood functions and so we can provide a cap of \( \alpha_\theta \) on the error of the estimate.

In practice, the use of stochastic models means that we are not able to make either of the assumptions, furthermore, the computationally complexity of the approach for large \( n \) deems
that we can’t practically compute small values of $\alpha_\theta$. While this theoretical approach does not have a practical implementation, it provides motivation to defining an measure on the suitability for a potential centrality and characteristic function.

With this in mind, we have ideal properties that we want for $C$ and $F_C$ to hold with respect to a parameter of the model. Furthermore, these properties allow for a logical measure of how close they are to being ideal. Formally we can introduce three measures of the properties:

1. The **max standard deviation** refers to the maximum standard deviation of all sample sets:
   \[
   M = \max (\{\sigma(F_C^\theta(G_\theta)) \text{ for } \theta \in \Theta\})
   \]  
   (13)

2. The **overlap of standard deviations** is the total length of overlap between the standard deviation intervals of neighbouring sample sets:
   \[
   O = \sum_k |\psi(F_C^\theta(G_{\theta_{\min}+k\alpha_\theta})) \cap \psi(F_C^\theta(G_{\theta_{\min}+(k+1)\alpha_\theta}))|
   \]  
   (14)

   where $\psi(S) = [\mu(S) - \sigma(S), \mu(S) + \sigma(S)]$.

3. The **continuity of standard deviations** is defined as the sum of difference between the standard deviation intervals of neighbouring sample sets:
   \[
   C = \sum_{k=1}^{n_s-1} \phi(\psi(F_C^\theta(G_{\theta_{\min}+k\alpha_\theta})), \psi(F_C^\theta(G_{\theta_{\min}+(k+1)\alpha_\theta})))
   \]  
   (15)

These provide a measure to compare against the ideal case which will have $M \rightarrow 0^+$, $O \rightarrow 0^+$ and $C \rightarrow 0^+$. Of course, the measure is susceptible to artificial reduction by reducing $s$ or increasing $\alpha_\theta$, as well as, the range of a centrality measure. For instance as shown in Figure 13, the range of degrees is a much smaller than the range of the characteristic path lengths. We can then provide some normalization with respect to these, but given the goal of these measures is not an exact evaluation, we will instead keep these limitations in mind when we are considering potential centralities. So we provide a simple normalization of these factors:

\[
E = \frac{\sqrt{M^2 + O^2 + C^2}}{R_s\alpha_\theta^{-1}} \rightarrow 0^+
\]  
(16)

where $R$ is the difference between the maximum and minimum of all sampled characteristic values.

Having access to a tool that will allow for a quicker initial consideration of potential centralities and characteristic functions provides similar functionality to that of some hypothesis tests in statistics. While it is not capable of determining exactly how the centrality would perform when evaluated, it is a heuristic to determine which centralities exhibit very poor properties and we can avoid the computational expense to fully evaluate it. Further, much of the computation is a subset of the computation of fully constructing a SL function and can be used without re-computation for the SL function so it is a ‘free’ screening.
6 Results

Having described motivation and reasoning behind our methodology, we are then interested to applying the method in a concrete manner onto as many centralities as is computational feasible. To do so we filter through proposed centralities with the screening methodology and then consider the filtered with the full evaluation methodology. We present which of the centralities have demonstrated sensitivity to parameter changes and provide low error estimates indicative of relevance to the measure of vascular structure. Further we demonstrate further application of the method to the analysis of real samples and the evaluation of other computational models.

When using the GBJ model for investigation we will use the following parameters as the default parameters. It is desirable to find default parameters and a feasible range that are ‘realistic’ and do not express the extremes of its behaviour. This will ensure that any particular default parameter does not overwhelm the other parameters in their influence of the vascular structure. For example, if we set our default $P_{branch} = 0.5$, then the extreme behaviour will increase the number of branches such that the $\theta_b$ parameter’s influence in the number of branches is not proportional and is not detected. The choice of these parameters has been done through experimental and visual inspection. With this in mind we enumerate the ranges and default values in Table 1. The other parameters are not varied and are summarized as:

$$diam = \frac{1}{200}, \sigma_{diam} = \frac{1}{10}diam, \sigma_v = \frac{1}{20}v_0, \sigma_\theta = \frac{1}{10}\theta_{branch}, AF_{x,y} = [(0.2, 1), (0.8, 1)].$$

We also have that unless otherwise noted, for all measures we use the interpretation of a graph that contains weighted edges denoting the number of cells that vessel traversed between branch points (length of the vessel), and, that the graph is undirected. Section 6.4 addresses the use of directed graphs.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default ($a_0$)</th>
<th>Range (Θ)</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Dist</td>
<td>5</td>
<td>[2, 8]</td>
<td>Number of cells</td>
</tr>
<tr>
<td>$P_{branch}$</td>
<td>0.2</td>
<td>[0, 0.5]</td>
<td>Probability</td>
</tr>
<tr>
<td>$\theta_b$</td>
<td>$\frac{\pi}{2}$</td>
<td>[$\frac{\pi}{6}$, $\frac{\pi}{2}$]</td>
<td>Internal Angle</td>
</tr>
<tr>
<td>$J_{dist}$</td>
<td>2</td>
<td>[1, 5]</td>
<td>Number of cells</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.5</td>
<td>[0.1, 0.9]</td>
<td>Ratio of AF gradient</td>
</tr>
<tr>
<td>$v_0$</td>
<td>1</td>
<td>[0.5, 1.5]</td>
<td>Ratio of N-Dist</td>
</tr>
</tbody>
</table>

Table 1: Default values of each parameter, $a_0$, and their corresponding ranges, Θ.

As has been reiterated, we have an initial interest in analysis of the small-world due to its prominent use and we are able to use the analysis as a benchmark for other considered centralities. We can then compare and contrast this with the measure of efficiency as motivated by the potential for the capabilities for the interpretation of the circulatory as a transport network. From here we present the analysis of other centralities that were introduced in Section 2.3. Having concrete results allows for the further discussion of which centralities are poised to provide the most relevance to the measure of structure similarity of vascular networks, or as a filtering centrality. Additionally, this provides a way to highlight the limitations and assumptions that are required in the methodology.

6.1 Analysis of Small-World Properties

Let us first establish a benchmark of measurements by using the small-world properties. This section will also be more methodical in the analysis of each step, this will allow for the enumeration of all considerations that are required in the evaluation methodology. Then we can refer to this template for the other evaluations of the centralities.

Described in Section 2.3.1, the minimum small-world properties are described as having a small mean characteristic path length (CPL) and a high global clustering coefficient. As is discussed in Section 6.1.2 we will decide to use the local clustering coefficient as we are able to use this to retain additional information in the structure of the network. As such we will consider these two centrality measures to represent the relevance. We note that the scale-free property will be discussed later as a part of the regularity centrality.

To demonstrate the use of the screening measures, let us initially consider the respective measures for the two centralities and interpret the values before we evaluate the performance of the parameter estimation. The measures are recorded in Table 2, and from the tables we can note the following:
Characteristic Path Length: We can guess that relative to the other parameters the error of \(N\text{-Dist}\) and \(J_{\text{dist}}\) will be similar and very small. Then the error of \(P_{\text{branch}}\) and \(v_0\) will also be relatively small while the errors of \(\lambda\) and \(\theta_{\text{branch}}\) will be large.

Local Clustering Coefficient: Contrasting with the CPL values, we see a larger value for all but the branching angle \(\theta_{\text{branch}}\). Hence, we will expect to see a better estimate of \(\theta_{\text{branch}}\) but have all other estimates be worse.

Then we will reference back here to see if these measures display a correct estimation of each parameter relative to one another. Hopefully, providing some validation in the heuristic that these measures provide for the relevance of the centralities. With this in mind, let us analyze the performance of the characteristic path length and local clustering coefficient in estimating the model parameters.

<table>
<thead>
<tr>
<th>Characteristic Path Length</th>
<th>Local Clustering Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N\text{-Dist})</td>
<td>(M) 13.8762 (O) 9.1811 (C) 0.9725 (E) 0.0296</td>
</tr>
<tr>
<td>(P_{\text{branch}})</td>
<td>(M) 131.9040 (O) 52.346 (C) 10.6025 (E) 0.0409</td>
</tr>
<tr>
<td>(\theta_{\text{branch}})</td>
<td>(M) 124.4670 (O) 9.9929 (C) 45.4009 (E) 0.0287</td>
</tr>
<tr>
<td>(J_{\text{dist}})</td>
<td>(M) 15.5720 (O) 5.6922 (C) 1.2091 (E) 0.0296</td>
</tr>
<tr>
<td>(\lambda)</td>
<td>(M) 126.7110 (O) 9.32222 (C) 47.2711 (E) 0.0292</td>
</tr>
</tbody>
</table>

Table 2: The screening measures of the characteristic path length and local clustering coefficient.

6.1.1 Characteristic Path Length

As is noted in the methodology, we must first inspect the distributions of the underlying CPL centralities to determine if we can use the proposed characteristic function. As denoted in Figure 15, we are able to see that the underlying distributions follow the expected normal distributions for the upper quantiles. The lower quantiles do not follow very closely and deviate upwards. Suggesting that there is a positive skew to the distribution, but is not an extreme skew. We could address this by applying a power transformation to shift the skew such that it more closely follows a normal deviation, but due to the claimed robustness of the method, let us first consider the measure without a transformation and apply if necessary.

![QQ-Plot of the CPL centrality distributions. The linear line denotes the expected normal distribution given the mean and standard deviation of the centralities. Each distribution is a collection of all the CPL values over 10 samples at the given \(P_{\text{branch}}\) values.](image)

Figure 15: QQ-Plot of the CPL centrality distributions. The linear line denotes the expected normal distribution given the mean and standard deviation of the centralities. Each distribution is a collection of all the CPL values over 10 samples at the given \(P_{\text{branch}}\) values.

Having defined our centrality, \(C\), and characteristic function, \(F_C\), we are able to construct the synthetic likelihood functions by generating the required samples over the interval of \(P_{\text{branch}}\). For the sake of demonstration let us first only consider the parameter \(\theta = P_{\text{branch}} \in \Theta\). Recalling that \(\theta_{\text{min}} = 0, \theta_{\text{max}} = 0.5, \alpha_{\theta} = \frac{1}{10}\) and \(N = 10\). So for each \(L_\theta(\theta)\) we compute Equations 6, 7, 3, 4 and 5 as the method describes. Then we have constructed 40 log SL functions at each value of \(\theta \in \Theta\), which we evaluate with a set of 100 test samples that are
generated, \{s\}. Figure 16 then plots the value of \( L_s(\theta) \) for some samples of the test data and we can compute the corresponding range estimate error by Equation 9:

\[
\frac{1}{|\Theta|} \sum_s |\theta_s - \hat{\theta}_s| \times 100 = (2)0.0150 \times 100\% = 3.0\%.
\]

We have that \( \alpha(\theta_{\text{max}} - \theta_{\text{min}}) = \frac{1}{10} \frac{1}{2} = 0.0125 \) and so the mean error is almost within a point of the computed granularity, while we don’t have other results to compare this with yet, it appears to be rather impressive in its capabilities of estimating \( P_{\text{branch}} \).

With reference to Figure 16, we can see that all the SL functions exhibit much of the positive behaviours described in Section 5.1.4. We are able to see that the functions are all uni-modal and the function is robust around the maximum points for the values of the extrema values of \( P_{\text{branch}} \), anecdotally these also resulted good estimations. We see that for the yellow plot of Figure 16a is our largest error and it corresponds to the plot having a decreased robustness as it has a similar shape to that of the step function, \( L_1 \). These attributes appear more common for the middle values of \( P_{\text{branch}} \). Overall, this does indicate that the positive behaviours in an SL function described corresponded to a good estimate and the others that were less robust had an increased error. In terms of the CPL, we see that it provided both a seemingly impressive error and SL function behaviour.

![Figure 16: Log synthetic likelihood functions. Each coloured line displays the values of \( L_s(\theta) \) over \( \Theta \). The lines correspond to a value of \( s \) and are selected from the test values such that they are spread over the interval. The crosses with the same colour represent \( \theta_s \) and the x-crosses denote \( \hat{\theta}_s \) from the corresponding \( L_s(\theta) \), the error estimate is the difference between these points.](image)

With this in mind, let us then consider the same process for the \( \theta_b \) parameter. Let \( \Theta = \left[ \frac{\pi}{6}, \frac{\pi}{2} \right] \), then we have the corresponding plot of the SL functions are given in Figure 17 and the range percentage error is computed as:

\[
\frac{1}{|\Theta|} \sum_s |\theta_s - \hat{\theta}_s| \times 100 = \frac{3}{\pi} 0.1364 \times 100\% = 13.0\%.
\]

When we consider the SL function plots in Figure 17, we can also see the stark contrast in the behaviours of the plots. Namely, we see that all the plots exhibit a combination of low robustness and multi-modal peaks in ranges around their parameter estimates. So it is capable of determining a region around the global maximum, but it is not decisive at the point of estimate due to a combination of these negative behaviours. Hence, these larger errors would account for the increased range error percentage with respect to the \( P_{\text{branch}} \) estimates.
The other 4 variable parameters are then summarized by their SL function plots and their range error percentages in Figure 18 and Table 3 respectively. Summarizing the behaviour here, we see that $v_0$ displays similar behaviour to $P_{branch}$ with regard to the range error percentage and the general behaviour of the SL functions. $\lambda$ and $J_{dist}$ both have similar inaccurate range error percentages, and, the sample plots of their SL functions denote that they provide little discrepancy of values over the range of the parameters, exemplifying both negative described behaviours. We note that these are relatively close to our threshold of 33% of guessing randomly and so they minimally detect any change in the network structure. We are to determine if this is due to the centrality being insensitive or if the parameters do not result in much change. It is also the case that the estimates of $N$-$Dist$ are extremely accurate and the SL functions are equally promising. However, it is natural to be skeptical of such results being related to the centrality measures, iterating only over 7 possible values and each change providing a relatively large deviation in the value would indicate that it is a property of the parameter. As such, we have noted the capabilities for the CPL and $F_C$ to estimate the parameters. However, we are interested to determine if this is a property of the CPL and $F_C$ of being specifically sensitive to these network changes, or if these changes in parameters are detected similarity by all other centralities. Further, given the accuracy of the error for $N$-$Dist$, $P_{branch}$ and $v_0$, we can validate the choice to not transform the underlying centrality distribution to better approximate the normal distribution.

Figure 18: Normalized log synthetic likelihood functions for the denoted variable parameters of the characteristic path length centrality. Refer to Figure 16 for description of plots.
Figure 19: QQ-Plot for the distribution of the local clustering coefficient. Each plot takes the values of all nodes over all samples, \(\{s_i^*\}\), of varied \(P_{\text{branch}}\). The linear line denotes the expected behaviour of a normal distribution.

![Unfiltered local clustering coefficient distribution](image1)

![Local clustering coefficient distribution filtered of zeros](image2)

![Mean of local clustering coefficient distribution over samples](image3)

![Standard Deviation of local clustering coefficient distribution over samples](image4)

(a) Unfiltered local clustering coefficient distribution.
(b) Local clustering coefficient distribution filtered of zeros.
(c) Mean of local clustering coefficient distribution over samples.
(d) Standard Deviation of local clustering coefficient distribution over samples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate Error</th>
<th>Range Error Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N-\text{Dist})</td>
<td>0.16</td>
<td>2.6%</td>
</tr>
<tr>
<td>(P_{\text{branch}})</td>
<td>0.0150</td>
<td>3.0%</td>
</tr>
<tr>
<td>(\theta_b)</td>
<td>0.1364</td>
<td>13.0%</td>
</tr>
<tr>
<td>(J_{\text{dist}})</td>
<td>0.0063</td>
<td>24%</td>
</tr>
<tr>
<td>(\lambda)</td>
<td>0.2135</td>
<td>21.4%</td>
</tr>
<tr>
<td>(v_0)</td>
<td>0.0434</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

Table 3: Range error estimates of the denoted model parameters. This computes the error of the estimates, \(|\theta - \hat{\theta}|\) over the test data, as well as, the error in terms of a percentage of the length of the parameter range, \(\Theta_\theta\).

### 6.1.2 Local Clustering Coefficient

The above process is then repeated with the local clustering coefficient centrality. We also noted that there are two ways to compute the global clustering coefficient, by counting the triangles or by taking the mean of the local clustering coefficient. Given we could retain more information about the network structure by having the local clustering coefficient for each node, we would prefer to use this measure. However, this centrality has an extreme positive skew of nodes with a value of 0, so much so that the median value is 0, denoted in Figure 19a. As a result, this would cause \(\Sigma_\theta^{-1}\) to be singular. We could then modify our centrality such that it filters out all values that are 0, which does improve the normality distribution from Figure 19b. However when we consider the mean and standard deviation of the unfiltered measures for all the samples of \(P_{\text{branch}}\), Figures 19c, 19d respectively, we can see that it does reasonably approximate a normal distribution, with the exception of an excess of zero values. However, these values are all computed when \(P_{\text{branch}} < 0.1\), and so we can use a modified characteristic function of

\[F_C(G) = \{\mu, \sigma\}.\]

We exempt the use of the polynomial regression as it would not provide a suitable fit to the step-wise function nature of the distribution.

With reference to Table 4 we see that the range error percentages have all substantially
increased aside from a slightly better estimate of $\theta_b$. In particular, the parameters estimates of $P_{\text{branch}}$ and $v_0$ have doubled, however they remain well below the threshold of 33%. Hence, they are capable of detecting some change in the underlying structure but we would assume that the precision is significantly worse around the estimates. This is highlighted by the SL function behaviour in Figure [20] we can see that there is particular issue with values of $P_{\text{branch}} < 0.3$ and that all the values here are overshadowed by greater parameter values. Even for the higher values of $P_{\text{branch}}$ we see the negative behaviours of many modes and a reduced robustness. These again account for the increased error in the parameter estimates.

When we consider $v_0$ we do have a more dampened behaviour with respect to the multi-modal functions, and this is reflected as the errors are not as significantly bad. We see that $\lambda$ and $J_{\text{dist}}$ have a large error in both of the measures, notably they are around or greater than a range error percentage of 33%. Similarly, the CPL was only able to significantly reduce the error from 33% to 24% and 21.4%, respectively. We note that if we were to generate an estimate of the parameter by drawing it uniformly from $\Theta$, the expected range percentage error would be 33%. Hence, for both measures the centralities are only marginally better than a random guess and are hence not sensitive to the parameters. Finally if we consider $\theta_b$, we see that the error is slightly better for the local clustering coefficient measure, and the behaviour is similar to as described for Figure[17].

In summary from investigation of the range percentage errors and the SL functions, we have seen that the positive behaviours described resulted in a reduced error and the negative behaviours resulted in an increased behaviour. As such, it may be the case that a centrality is good at estimating a parameter in a particular sub-range and we can use the SL function behaviours to determine which part of the range the centrality is better suited for parameter estimation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate Error</th>
<th>Range Error Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$-Dist</td>
<td>0.6</td>
<td>8.6%</td>
</tr>
<tr>
<td>$P_{\text{branch}}$</td>
<td>0.1026</td>
<td>20.5%</td>
</tr>
<tr>
<td>$\theta_b$</td>
<td>0.1329</td>
<td>12.6%</td>
</tr>
<tr>
<td>$J_{\text{dist}}$</td>
<td>0.0076</td>
<td>38.0%</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.3006</td>
<td>37.6%</td>
</tr>
<tr>
<td>$v_0$</td>
<td>0.1108</td>
<td>11.1%</td>
</tr>
</tbody>
</table>

Table 4: Range error estimates for the local clustering coefficient, denoted by model parameters. For further description refer to Table 3.

From this we can also note that our results were similar to what the screening measures in Table 2 had estimated of the two centralities, namely that we would have an inferior measure for all the parameters except for $\theta_b$. As another side note we can note the run-time of constructing the SL functions as:

$$\approx \frac{112}{N}\text{-Dist} + \frac{196}{P_{\text{branch}}} + \frac{181}{\theta_b} + \frac{25}{J_{\text{dist}}} + \frac{192}{\lambda} + \frac{364}{v_0} = 1070 \text{ seconds.}$$

This simply allows for a point of reference of computation time for $N = 10$ and was found to be generally representative of most centralities. Having established that there is difference
in the capabilities for the two centralities to estimate the values of parameters. Noticeably, the estimates for $P_{\text{branch}}$ and $v_0$ produced a large discrepancy in the estimated errors. We then want to explore other possible centralities to determine if we can find more appropriate measures to recommend. Additionally, we have provided a thorough demonstration of the steps required to evaluate a centrality within the methodology, as such, we will follow the same process for the following mentioned centralities. However we will omit descriptions for many of the steps that are repetitive in the process to allow for discussion, so we can assume that this process was followed and any differentiation will be noted.

### 6.2 The Four-Dimension Approach for Potential Centralities

Having established benchmark measures with respect to the CPL and local clustering coefficient, we can explore how the estimations perform for other potential centralities. As we described in Section 2.3.2, we want to consider centralities that belong to the different categorizations of the four-dimension approach. We will then structure the following results by investigating the centralities in the same order as they were previously described.

#### 6.2.1 Reachability

Recalling that, in the context of a vascular network as a transport network, the distance measures denote a measure of how quickly a node could transport nutrients to another node. Given that the efficiency is the reciprocal measure of the CPL, we would expect to have similar error estimations, however, with reference to Table 5, we see that the CPL is inferior with respect to the efficiency in all measurements. As was mentioned, we might attribute this to the loss of information that is given when the graph is not connected and having been required to modify our computation to be component-wise, whereas, the efficiency is able to mitigate this issue by having it in the denominator. Hence, we favour the efficiency centrality.

In contrast, the degree centrality is less important on the individual node level and is rather the distribution of the degree that is of interest. Interestingly, the simple measure of degree centrality provides similar results in parameter estimation to the efficiency but the more intricate measures degrade in performance. Since the Katz and eigen vector centrality consider a greater neighbourhood around the node, these values would become more homogenized throughout the distribution of the network reducing the range in the distribution and making it less sensitive. The general behaviour of the SL functions of the efficiency and degree follow similar behaviours to those in Figures 16, 17 and 18, so we refrain from providing the redundant figures and similar conclusions can be drawn. From this we want to further consider the efficiency and degree centrality when constructing a centrality.

<table>
<thead>
<tr>
<th>N-Dist</th>
<th>CPL</th>
<th>Efficiency</th>
<th>Degree</th>
<th>Katz</th>
<th>Eigen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2400</td>
<td>4.0%</td>
<td>1.7%</td>
<td>0.0100</td>
<td>0.2%</td>
</tr>
<tr>
<td>$P_{\text{branch}}$</td>
<td>0.0186</td>
<td>3.7%</td>
<td>0.1000</td>
<td>3.0%</td>
<td>0.0149</td>
</tr>
<tr>
<td>$\theta_b$</td>
<td>0.1312</td>
<td>12.5%</td>
<td>0.1416</td>
<td>13.5%</td>
<td>0.1393</td>
</tr>
<tr>
<td>$J_{\text{dist}}$</td>
<td>0.0086</td>
<td>42.8%</td>
<td>0.0041</td>
<td>20.5%</td>
<td>0.0076</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.2303</td>
<td>28.8%</td>
<td>0.2205</td>
<td>27.6%</td>
<td>0.2164</td>
</tr>
<tr>
<td>$v_0$</td>
<td>0.0530</td>
<td>5.3%</td>
<td>0.0423</td>
<td>4.2%</td>
<td>0.0437</td>
</tr>
</tbody>
</table>

Table 5: Estimate and range percentage errors for reachability centralities, represented in the first and second column of each centrality respectively.

#### 6.2.2 Amount of Flow

The stress and betweenness centralities consist of those that measure flow through nodes in the network. However, when attempting to construction the SL functions of the stress centrality we have issue with $\Sigma_{\theta}$ being singular, this is thought to be due to a loss of precision due to the very large numbers of the stress values. Due to the normalization in the betweenness centrality this is not the case and so we are only able to consider it as our measure of flow, shown in Table 6.

37
In comparison with the other presented measures so far we see that it is worse but only marginally so. Given the physical connotations of the flow within a vascular network, we would hope that this would provide a better measure of the network. Particularly when we consider the many biological mechanisms of the blood vessel to have control over the blood pressure and vessel dilation in order to control blood flow. However, we note here that this is the betweenness centrality without incorporating the directional flow of a directed graph or by providing additional weights to the edges in order to denote the diameter of each blood vessel. To consider this biological motivation we discuss this further and explore incorporation the directed measure in sub-section 6.4. As such, this measure would currently be overshadowed by the previous measures discussed.

6.2.3 Vitality

When we first considered, we were unfortunately forced to dismiss the original measure due to the computational infeasibility of re-computing the centrality for each node in a network, instead we introduced the measure of robustness with respect to the radius and diameter measures.

Given the hierarchical shape of the vascular network as shown in the sample networks of Figure 11, it is apparent that only in cases when there are very few nodes that we would have that removal of a particular node would segment the network into more components and increase any shortest-path lengths. Further, in the case when there are few nodes, we have that the shortest-path is already either infinite among components and creating another isolated node would not differ the measure noticeably. Hence, the generated vascular networks have two phases of being highly redundant when in a dense hierarchy or having all nodes be vital when there are relatively few of them. However, this measure is unable to distinguish between these two states and so we have the unappealing results shown in Table 7 in our parameter estimates that are about as good as guessing the parameter uniformly on the entire range. So we will not consider these measures in our candidates for measuring similarity.

<table>
<thead>
<tr>
<th>Centrality</th>
<th>Radius</th>
<th>Diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Dist</td>
<td>2.1500</td>
<td>1.2667</td>
</tr>
<tr>
<td>P_branch</td>
<td>0.1855</td>
<td>0.2110</td>
</tr>
<tr>
<td>θ_b</td>
<td>0.4497</td>
<td>0.3821</td>
</tr>
<tr>
<td>J_dist</td>
<td>0.0053</td>
<td>0.0067</td>
</tr>
<tr>
<td>λ</td>
<td>0.1861</td>
<td>0.1949</td>
</tr>
<tr>
<td>v_0</td>
<td>0.4081</td>
<td>0.2866</td>
</tr>
</tbody>
</table>

Table 7: Estimate and range percentage errors for measures of robustness (vitality) centralities, represented in the first and second column of each centrality respectively.

6.2.4 Feedback

A feedback centrality is one that uses an underlying centrality, C, and their values. In particular we focused on the biologically motivated centralities that provided measures on the degree distribution of the network, focusing on the expected regularity of the generated networks. The measures of assortativity, regularity and entropy were proposed as suitable centralities to consider this.
Table 8 denotes the error values, in which we note that the regularity has lost precision of the degree centrality for all of the relevant parameters. We would expect this as the measure of regularity, should stay close to the normalized value as the network grows. In contrast, the degree distribution would then be more sensitive to changes in the network growth and hence hold better parameter estimates. What is interesting is that the increase in error is less that was initially suspected, given we hold this assumption of a highly regular network. However, this could be due to the implementation of the network model that allows for nodes with a degree greater than 3 to help simply the modelling process. With this in mind, we can disregard the centrality in favour of the regular degree centrality.

As we have described, the presumed regular nature of the vascular networks would imply that the graph would have rather low entropy and that it would be consistently. However, as the error highlights, the entropy provides results that are close to the established centralities so far. With reference to Figure 22, we see that the entropy does not converge in a similar manner to the assortativity and rather it is distinguished for $P_{\text{branch}}$ and $v_0$. Given the global nature of the measure, it is impressive that is almost equivalent with the values of the mentioned centralities, and suggests that the entropy is particularly well suited for the parameter of $P_{\text{branch}}$. As such, we can recommend its use for this parameter.
Figure 22: Sample entropy values over the ranges of their respective parameters. Each parameter has a granularity of $\frac{1}{40}$ and a sample network at each granularity point is taken.

6.2.5 Summary

Having considered the mentioned centralities that fall under our different basic term categorizations, we see a distinction of the reachability and flow centralities as being used for measuring the difference in networks. In particular, we note the centralities of the efficiency, degree and betweenness show promise as a similarity measure of difference between vascular networks. It was also shown that the measures of robustness provide no information on the differentiation between network structures. When we considered the feedback measures and the biologically motivated reasoning, it was shown that the assortativity allows for a filter for a range of potential vascular networks. Finally, we also considered the entropy measure which proved to be surprising in its accuracy of estimating the $P_{branch}$ parameter.

With this in mind, we want to consider potential sources of invalidity that could be argued for the measure of the centralities. In the following sections we address potential arguments as to the error estimates not being a property of the centralities for these established centralities.

6.3 Characteristic Function Implications

Noticeable was the overall increased error when we considered the feedback and the vitality measures. While we can attribute these to the biological implications that were mentioned, we also note the common trait that they all provided a global measure of the network and so their characteristic function was reduced to being the identity function. In this manner, we no longer summarize the entire distribution of the nodes throughout the network and appear to be losing information. As such, we want to ensure that the centralities that have provided good estimates are not a property of the characteristic function and rather a property of the centrality.

<table>
<thead>
<tr>
<th></th>
<th>Mean Efficiency</th>
<th>Mean Degree</th>
<th>Mean Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$N$-Dist</strong></td>
<td>0.3000</td>
<td>4.9 (+3.2)%</td>
<td>0.0200</td>
</tr>
<tr>
<td><strong>$P_{branch}$</strong></td>
<td>0.0344</td>
<td>6.8 (+3.8)%</td>
<td>0.0123</td>
</tr>
<tr>
<td><strong>$\theta_b$</strong></td>
<td>0.1368</td>
<td>13.3 (+0.3)%</td>
<td>0.2100</td>
</tr>
<tr>
<td><strong>$J_{dist}$</strong></td>
<td>0.0056</td>
<td>27.9 (-0.3)%</td>
<td>0.0082</td>
</tr>
<tr>
<td><strong>$\lambda$</strong></td>
<td>0.2152</td>
<td>26.9 (-8.9)%</td>
<td>0.2521</td>
</tr>
<tr>
<td><strong>$v_0$</strong></td>
<td>0.0728</td>
<td>7.2 (+3.0)%</td>
<td>0.0491</td>
</tr>
</tbody>
</table>

Table 9: Estimate and range percentage errors for the centralities using the simplified mean characteristic function, represented in the first and second column of each centrality respectively. The relative increase or decrease in the range percentage error is denoted by the colour-coded percentages.

Table\[^9\] denotes the errors retained for the efficiency, degree and betweenness centralities with a characteristic function that only outputs the mean value. From inspection of the table we see that the difference in characteristic function does contribute to the accuracy of error estimation. Other than the 4 green highlighted cases, we see that by reducing the information retained in our characteristic function the estimation precision decreases. The 4 cases were within $J_{dist}$ and $\lambda$ when the estimates were already close to the 33% threshold of randomness guessing. This provides evidence that by using the characteristic function, the information is better retained and used to improve estimations. Furthermore, we see that with the exception of the mean betweenness measure of $P_{branch}$ that the difference is not so
substantial as to make the measures equivalent to the vitality and feedback measures. Hence, we want to use the characteristic function with the maximum information when capable and the difference in the error estimates appear to be a property of the efficiency, degree and betweenness centralities rather than solely of the characteristic function. This reinforces the previous assertions that we have made about each of the centralities.

6.4 Incorporating Directional Measures

In the same vein of considering how the retention of information allows for better error estimation, particular for the context of our measure of flow through the network from the physical interpretations of the measure. Given the additional information allows for a better representation of the network structure, we would expect that the measures are capable of providing better estimates. In this manner we compare the performance for the given centralities of the efficiency and betweenness.

<table>
<thead>
<tr>
<th>N-Dist</th>
<th>Directed Efficiency</th>
<th>Directed Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0900</td>
<td>1.5 (-0.2)%</td>
<td>0.1100 1.8 (-0.4)%</td>
</tr>
<tr>
<td>P_{branch}</td>
<td>0.0161</td>
<td>3.2 (+0.2)%</td>
</tr>
<tr>
<td>θ_b</td>
<td>0.1482</td>
<td>14.1 (+0.6)%</td>
</tr>
<tr>
<td>J_{dist}</td>
<td>0.0095</td>
<td>47.5 (+27.0)%</td>
</tr>
<tr>
<td>λ_v_0</td>
<td>0.2221</td>
<td>22.2 (-5.4)%</td>
</tr>
</tbody>
</table>

Table 10: Estimate and range percentage errors for the centralities when the measures are altered to use the directed graph representations of the networks. The relative increase or decrease in the range percentage error is denoted by the colour-coded percentages.

We would prefer centralities that are accurate without requiring the additional knowledge of blood flow due to the increased complexity of determining this from real world samples. While it is easily found from most computational models, if we want to compare the model with a real world sample and the measure is dependent on such information we are not able to extract then we reduce the applicability of the method. As shown in Table 10 the lack of any significant improvement, when using the directed versions of the networks lets us continue using the undirected networks without worry that the information is vital to the performance of the measures. Hence allowing the methodology and suggested measures to be generally more applicable.

6.5 Assumption of Sample Sizes

As has been of focus, we need to make the assumption that our characteristic function values follow a normal distribution, and a key factor is the number of samples that we need to consider such that we can use the central limit theorem. As was introduced, we have used 10 sample at each of 40 granularity points for the parameters of P_{branch}, θ_b and v_0. This was originally chosen as it appeared to provide similar results to higher sample sizes and was much more reasonable in computation time. However, it is of course important to investigate that this is indeed the case and that we are not losing accuracy due to reduced sample size.

<table>
<thead>
<tr>
<th>Efficiency</th>
<th>Degree</th>
<th>Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_{branch}</td>
<td>θ_b</td>
<td>v_0</td>
</tr>
<tr>
<td>μ</td>
<td>0.0128</td>
<td>0.1364 0.0435</td>
</tr>
<tr>
<td>σ</td>
<td>0.0015</td>
<td>0.0035 0.0015</td>
</tr>
<tr>
<td>Δ_1</td>
<td>0.2164</td>
<td>0.4094 0.4312</td>
</tr>
<tr>
<td>Δ_5</td>
<td>0.0023</td>
<td>0.0093 0.0051</td>
</tr>
<tr>
<td>Δ_10</td>
<td>0.0041</td>
<td>0.0056 0.0007</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Efficiency</th>
<th>Degree</th>
<th>Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_{branch}</td>
<td>θ_b</td>
<td>v_0</td>
</tr>
<tr>
<td>μ</td>
<td>0.0148</td>
<td>0.1353 0.0415</td>
</tr>
<tr>
<td>σ</td>
<td>0.0003</td>
<td>0.0062 0.0018</td>
</tr>
<tr>
<td>Δ_1</td>
<td>0.2193</td>
<td>0.4110 0.4285</td>
</tr>
<tr>
<td>Δ_5</td>
<td>0.0026</td>
<td>0.0257 0.0035</td>
</tr>
<tr>
<td>Δ_10</td>
<td>0.0008</td>
<td>0.0100 0.0022</td>
</tr>
</tbody>
</table>

Table 11: For each centrality, we follow the framework differentiating the number of samples, N, taken at each point of θ ∈ Θ, where N ∈ {5, 10, 15, 20, 25, 30, 35, 40, 45}. Then we measure the error for each one over the same 100 test samples. μ and σ denotes the average and standard deviation of the error over all N, respectively. And Δn denotes the difference between when N = n and the minimum error achieved over all N.

As such we show by the standard deviation values in Table 11 that the range of errors over N is extremely small and that the accuracy has converged. Furthermore, by the Δ_{10}
measures it appears that the error has already converged to the minimum for $N = 10$ and so we are justified in our choice of $N = 10$. In fact, it appears that $N = 5$ would have also been suitable as it was only marginally worse than $N = 10$ but would be approximately half of the wall run time. Having noticed this, we can then also check the performance when $N = 1$ in the hopes of further reducing the computational requirements. As is shown in the table however, we see that the error unanimously are about the same or worse than guessing randomly. As such, we can recommend the use of $N \geq 5$, or equivalently about 200 total samples, with strong diminishing returns with respect to the computational wall time. Such behaviour validates the presented errors with $N = 10$, and that we could have reduced $N = 5$ to achieve similar results with a reduced run-time.
7 Discussion

We recall the aim of the methodology, wanting to construct a method in which we could evaluate how well centralities are able to estimate the parameters of a model that generates tumour-induced vascular networks. A measure that has a minimal error across the parameters is sensitive to the underlying structural changes of the vascular networks and implies that it is a relevant measure to the comparison of them. Having established such a method, we then applied it to many centralities to allow for contrasting between them.

Motivated by the current use of the small-world properties for the evaluation of computational models of angiogenesis we used these as benchmarks. Then we wanted to provide further quantitative evidence that they should be used as such or we could demonstrate that there exists more proficient centralities. From the results, we have shown that indeed the characteristic path length, or its reciprocal measure of efficiency, provides one the most capable error estimation within the methodology. However, the global clustering coefficient did not provide the same proficiency. Since real vascular networks do exhibit a relatively high global clustering coefficient and as a bi-product of our methodology we have evaluated the global clustering coefficient on a wide variety of tumour-induced vascular networks, then it is instead proposed to use the measure as a filtering measure to determine if a network is a vascular network.

Along with the efficiency, the results indicate that the measures of the degree and betweenness centralities exhibit similar performance in the error estimation of the parameters. Hence, by the same reasoning these can also be recommended as measures of similarity for vascular networks. Additionally, we also wanted consider if a combination of the measures were capable of providing a better constructed centrality. Notably from the results, there were a number of concerns and considerations that should be discussed to ensure that the interpretation of the error estimates is correct. Let us enumerate through these in the following paragraphs.

The first point of issue is determining which parameters provide the most insight given the general ranges of the parameter estimation errors, since we can see from the results that the minimum achieved error for each parameter varies substantially. Determining the potential ranges would be an important step in being able to provide validity to if a centrality is sensitive to the structural network changes, or if the parameter simply does not induce significant change.

We can recall that if we simply guess a parameter within the parameter range, then our mean range percentage error would be 33%. Hence, if all centralities also have a similar average error rate, then it would imply this is because the generated networks do not significantly change from the parameter variations. This empirically appears to be true for $\lambda$ and to a lesser extent of $J_{dist}$. The minimum error for $\lambda$ across considered centralities was 23.3% in our radius robustness measures but overwhelmingly the measures were consistently over 26%. As such, it is considered that the parameter does not provide significant change to the network structure and should not be considered in our evaluation of centralities. Physically, $\lambda$ denotes the decay rate of the angiogenic factor and it only dictates in which direction the vessels should grow, but it does not change the rate of growth. However, if the growth factor is strong then it might be possible that it would cause the network to spread over the different sources and reduce its connectivity. Hence why the measure of robustness would have been sensitive to the changes.

In the same vein, $J_{dist}$ is the distance that tip cells need to be in proximity with one another to join and so physically, before even considering the results, changing this would appear to not create much structural change. In large part due to the fact that, if a vessel did not join in current growth phase, then it would still be highly likely to join in the next phase. This is largely reflected in the poor parameter estimations across centralities, and so for the physical reasons and given results it is also not considered as a good parameter to consider as a measure of the centralities.

In contrast, $N-Dist$ produces an accurate error estimate for most centralities. When we consider the parameter physically, this is not surprising as it controls how many cells a vessel will grow in each iteration of the GBJ algorithm. Due to the implementation requiring that this is a whole number, we can only iterate over its parameter range at a granularity of 1 and the range is [2, 8]. Hence, each perturbation would result in a very large structural change of the network and these positive error estimates can not be attributed to the centralities. Therefore we also cannot consider this parameter when evaluating the centralities.

The other parameters however, all exhibit a similar behaviour in their error estimates over the considered centralities, in which there is a wide range of potential errors. $P_{branch}$ being the best example as it has a bi-modal distribution of error estimates over the range
of $\approx 3 - 25\%$. This is indicative of error estimates being a property of the centrality rather than the parameter. This follows similarly for $\theta_b$ and $v_\theta$. Therefore, when we attribute which parameters provide the most insight, we mainly consider $P_{\text{branch}}$, $\theta_b$, and $v_\theta$.

Having differentiated between the error estimate being a property of the parameter or the centrality, we further need to consider if the error estimates could be attributed to the characteristic functions that were being used rather than the centralities. Some measures that we considered are a global measure of the network rather than a centrality in the traditional sense. After noticing that these measures performed significantly worse, it was required that we measure the other centralities with a characteristic function that would retain similar information to the global measures. As demonstrated in Section 6.3, we showed that while reducing the information in the characteristic function negatively impacted the performance of the measures, the performance of the centralities was still significantly better than the global measures mentioned. Hence, there is evidence that the characteristic function is important to help reduce error, but we can still attribute the error estimates to the corresponding centralities. Furthermore, we should recommend the use of characteristic functions that retain more information when measuring the similarity between networks.

In the same vein, we also wanted to consider if retaining physical information of directional flow in the network would allow for better error estimates. From the results it was evident that the impacts were marginally better, however, it is of note that the inclusion of this information can change what the centrality measures. For instance, when we compare the measure of efficiency on a directed vs undirected graph then any nodes that are downstream of a node will not have a shortest-path back upstream. As such, the number of nodes that have an efficiency close to 0 is increased substantially and this could cause a skew in the underlying centrality distribution, which jeopardizes our assumption of the normal multivariate distribution. However, it was shown that the distribution remains similar and we can continue to use the assumption of the multivariate distribution. Therefore, we have evidence that the additional information represented by the directional flow does not produce a noticeable increase to the accuracy in a similarity measure.

In further validation of the presented results, we also demonstrated how the error of the methodology converged quickly with respect to $N$, the number of samples taken at each granularity point, $\alpha_\theta$. In this manner, it was shown that when $N = 1$ the errors were so significant that it did not allow for any differentiation between centralities and that they were all about the same or worse than randomly guessing the model parameters. However, from $N \geq 5$ the errors had converged and given the linear increase in the computational time as $N$ increases, we have a scale of diminishing returns to increase $N$. We can also note that when $N = 5$ this results in 200 total samples over the entire range of $\theta$ which is about the same as $\sqrt{50000} \approx 223$ as was used in the original application of the synthetic likelihood function by Wood [24].

### 7.1 Limitations

As has been noted throughout, there are limitations in the proposed methodology primarily regarding the computational feasibility. This arises naturally from the desired sample sizes when we want to use various statistical methods. When we consider the initial implementation of the synthetic likelihood function by Wood [24], they use a Monte-Carlo method and recommend a sample size of 50 000 for a model with 2 parameters, or equivalently around 223 for each dimension of the parameters. Even in the simplified GBJ model we still have a high dimensionality of 6 parameters that we consider and the number of samples will need to grow exponentially to consider the same granularity of each parameter. This quickly becomes infeasible, recalling that when measuring the single parameters, we considered the evaluation of 400 samples which had an approximate average wall time of 1070 seconds [17]. As a consequence of having these infeasible computational requirements for increased dimensionality, we can not guess numerous parameters simultaneously as we had originally outlined in the methodology.

There are two approaches to help reduce the computational requirements that come to mind, centred around reducing the number of required samples. The first would be to establish a smaller range of the possible parameters to allow for a similar granularity in measuring. As such, we would require a survey or further research into a large range of vascular networks and determine what are more constricted ranges of the parameters given the type of vascular network considered. The other would be to decrease the granularity of the parameter ranges which we considered when discussing the impact of sample sizes on the network. As was noted however, we saw the degradation of the estimates if we reduced the number of samples below 400 on the interval. Further the average error is bounded by the granularity of the
estimates and so we are also bounded by the possible reduction. With these considerations, we see that the computational requirements for the methodology using an underlying model with an increased dimensionality is infeasible without strong restrictions on the parameter ranges.

When we consider the presented vitality measures, it was noted that there are also limitations when we consider a computationally demanding centrality measure. In the case of the radius and diameter measures, there were efforts to optimize the implementation of the centralities to only recompute the shortest path when the deleted node was in the path, however, such optimizations are proportionally irrelevant to the overall work required. While this is an extreme instance in which the total required work grows quadratically with respect to the graph size, it demonstrates that there are limitations on the considered centrality functions.

The other main limitation is with respect to the chosen GBJ model. As we noted, the model was chosen due to the simplicity it required and the scale at which the model functioned. This allows for a model that had much less dimensionality in the parameters than other models, which, as demonstrated, is a major constriction. The scale also allowed the construction of a more complete representation of a capillary bed in the vascular network which allows for a wide range of dynamical behaviour. However, these properties come at a trade-off to the accuracy or realism of the model as we are simplifying much of the complex behaviour and mechanisms of angiogenesis. This is addressed by the flexibility poised in the methodology, as we are easily able to use any suggested model to generate the sample networks.
8 Conclusion

Due to the detrimental implications of cancer, it has been the focus of research across many research domains with the goal to allow for the treatment of cancer. Of these different research domains, predictive medicine aims to provide ways to predict the growth of cancer to allow for preventative procedures. It is established that angiogenesis is a primary way for tumours to gain access to required nutrients from the vascular network of the host. Further, connecting to the vascular network allows the tumour cells to colonize distant organs and become even more threatening to the host. There is, thus, a need to develop an understanding of angiogenesis so that better methods can be applied to predictive medicine.

Research into creating computational models of angiogenesis, and more specifically tumour-induced angiogenesis, aims to provide this deeper understanding. Consequently, a large quantity of number of computational models have been and are being produced, however, from surveying the work done in this field, and in particular the evaluation of each model’s performance, there is a large dependence on using qualitative measures. Such measures are largely required to been done manually and a lack of quantitative measures make such comparisons even more difficult. This leaves a need for such measures to be established to allow for more accurate and automated evaluation of model performance.

To address this need, we can use network centralities as a quantitative measure of the structure present in a vascular network. Currently in the evaluation of computational models, the small-world properties are the primary, and almost only, centrality used as a quantitative metric. The motivation of using such a metric is largely limited to the reasoning that biological networks exhibit small-world properties and there are not centralities or properties that are more specified for vascular networks. With this in mind, this thesis describes a methodology to evaluate the relevance of centralities for use as such a quantitative measure. Throughout the results it is shown that the methodology provides distinction between the capabilities of centralities to estimate the parameters of a computational model. Namely, we have established that the methodology is sensitive to different choices of centralities, characteristic functions and the physical information that is retained in the network representation. This allows the conclusion that a centrality capable of minimal error estimation is sensitive to changes in a vascular network structure and is therefore relevant to a measure of similarity between networks. Conversely, if a centrality produces similar error estimates to that of a random guess, then it implies that all the generated vascular networks have a similar measure and the centrality can be recommended as a filter to determine if a network has properties of a vascular network.

Given the ability to discern relevance of centralities, we investigate the characteristic path length, global clustering coefficient and the degree distribution associated with the small-world properties. The results show that the characteristic path length provides minimal error estimates of the parameters and hence there is evidence that it should be used to measure the difference between vascular networks rather than filter out if it is a vascular network. Inversely, the global clustering coefficient did not provide sufficient estimates and is indicative to be used as a way to discern a vascular network. The degree distribution was decisively not scale-free, and so a measure of the regularity of the graph can be used in the same manner of the global clustering coefficient. However, the degree centrality performed similarly to the characteristic path length and the same conclusions can be drawn. Therefore, we provided further evidence that the small-world properties are relevant measures of a vascular network structure. It is recommended to use the global clustering coefficient and regularity as a measure of whether a network exhibits properties of a vascular network, and to use the characteristic path length as a measure of difference between vascular networks. Hence in the evaluation of a computational model with respect to a real network, it is not sufficient to denote it simply is a small-world network but that the characteristic path length is also similar to the network being emulated.

We then showed using the same arguments that the efficiency, degree and betweenness centrality are preferred over the characteristic path length due to providing minimal range percentages errors, although these were marginal improvements and so the characteristic path length can still be used. It was also established that the measures of robustness provide similar functionality as the global clustering coefficient and should also be used when considering if the network exhibits properties of a vascular network. As such we recommend that a vascular network should exhibit a high value of robustness with respect to the diameter or radius.

The establishment of the methodology and the well-performing centralities of the efficiency, betweenness and degree centrality allow for a specialized quantitative measure of how similar vascular networks are. This provides a tool for the development of future computational models as a way to measure the difference between the generated model and an example real network that is being emulated. Additionally, it opens the opportunity to review established
computational models and compare them with one another. In the context of predictive medicine, we have also provided relevant measures in which to contrast the structure of real networks. As such, we could use the measures to determine how a vascular network formed through normal and tumour-induced angiogenesis differ. Assuming such measures would provide a distinction between the types of networks, it would allow for the a measure to determine the likelihood that a vascular network is tumour-induced and allow for the screening of cancer.

In this manner, we can consider potential future work to be done that largely involve experimenting with the aforementioned use-cases. It would prove beneficial to use the methodology over a large number of computational models to allow for the ability to contrast the error estimates of parameters further. This would provide further insight into what error estimate ranges should be expected and how we can interpret them. As we mentioned in the limitations, we made a choice of the GBJ model due to its simplicity, however, there could be a better suited model that we can now discover having established the methodology and how to evaluate it.

As we have established recommended centralities for the comparison of the networks, it then will be required that we compile a collection of vascular networks extracted from samples of normal and tumour-induced vascular networks. After collecting such samples we would be able to follow the previously described steps to construct a predictive measure.

Further, it could also be possible to combine the recommended centralities into a combined single centrality using the four-dimension approach. This could be approached by constructing an optimization problem over the weights of the centralities. Given the large computational cost that such a optimization would cost it was not possible to consider in the context of this thesis, and it is also not immediately evident if such a measure would provide much convenience over considering the 3 centralities all together.
References


9 Appendix A - GitHub Repository

For the sake of reproducability, linked here is a GitHub repository that includes everything used to create the thesis. At the head of the directory, there exists pres.* files that include the content for the corresponding presentation of the thesis, thesis.* files for the source of this document, a figs folder containing all used figures, and a src folder containing all required code to generate the results given.

The julia code was written with the use of plotting and graph libraries [26, 27].

10 Appendix B - Growth-Branch-Join Implementation

For the GBJ model we use the work of Stefan Engblom with permission. The implementation of the model is listed below as MatLab code, with minor modifications.

```matlab
% TESTGBJ Test of a simple Grow–Branch–Join description of blood networks.
% The purpose here is to investigate if a rather simple probabilistic description of the growth of blood vessels can be used as an implicit definition of a likelihood over blood vessels, i.e., graphs in the unit square.
% S. Engblom 2023–01–27

%% (0) model parameters: either from input struct or using default values
visevent = false; % visualize events at all
visB = visevent & true; % visualise branching (green circles)
visJM = visevent & true; % visualise Join/Merge (blue diamonds)
visJMI = visevent & true; % visualise Join/Merge Into (black squares)
visAF = false; % visualize gradient field of AF
toggle_plot = true;
if toggle_plot
figure(1), clf, hold on,
end

if exist('par','var')
% input struct
seed = par.seed;
RAND = par.RAND;
diam = par.diam;
sigma_diam = RAND*par.sigma_diam;
NDdist = par.NDdist;
v0 = par.v0;
sigma0 = RAND*par.sigma0;
Pbranch = par.Pbranch;
theta = par.theta;
sigma_theta = RAND*par.sigma_theta;
Jdist = par.Jdist;
lam = par.lam;
AFx = par.AFx;
else
% default values:
seed = 8912345;
RAND = 1; % 0 or 1 to turn randomness on/off
% cells
diam = 1/200; % cell diameter
sigma_diam = RAND*0.1*diam; % std cell diam
NDdist = 5; % Notch–Delta distance (# cells)
```
% Growth
v0 = NDdist*diam;  % speed of growth at unit time steps
sigma0 = RAND*0.05*v0;  % directional noise

% Branch
Pbranch = 0.2;  % probability to branch
thetab = pi/6;  % branch angle
sigma_thetab = 0.1*thetab;  % std branch angle

% Join
Jdist = 2*diam;  % if within distance

% AF
lam = 0.5;  % decay of AF
AFx = [0.25];  % position of AFs (x,y = 1)
end

% (1) initializations

% random generators
randgn = RandStream.create('mlfg6331_64','seed',seed,...
    'NumStreams',4,...
    'StreamIndices',1:4,...
    'CellOutput',true);
% We consistently use randgn{1} for the initial positioning of the tipcells, randgn{2} for the branching angles, randgn{3} for the growth SDEs, and randgn{4} for selecting branching tipcells.
% fit the sum of NDdist i.i.d. lognormal diameters to "mean +/- std =
% diam +/- sigma_diam"
d2 = (NDdist*diam).^2;
s2 = (NDdist*sigma_diam).^2;
mu_branch = log(d2/sqrt(s2+d2));
sigma_branch = sqrt(log(s2/d2+1));

% the vessel geometry is the unit square, with AFs placed as unit % point sources at y = 1:
AFx = permute(AFx,[1 3 2]);  % (3rd dim used for the n sources)
AFy = ones(1,1,size(AFx,3));

% Green's function for the operator T(u) = Laplace(u) - lam*u in 2D:
Gfun = @(x,y,px,py,lam)=1/(2*pi)*... besselk(0,sqrt(lam)*abs(complex(x,y)-complex(px,py)));
% using BesselK0' = -BesselK1 and the 2D gradient expressed using % complex numbers
Gfungrad = @(x,y,px,py,lam)=1/(2*pi)*... (besselk(1,sqrt(lam)*abs(complex(x,y)-complex(px,py)))).*... sqrt(lam).*... (complex(x,y)-complex(px,py))./abs(complex(x,y)-complex(px,py));

% concentration of AF at (x,y) (note: solving u_t = Delta u - lam*u + f in % stationary form yields Delta u - lam*u = Tu = -f)
AFconc = @(x,y,lam)=sum(Gfun(x,y,AFx,AFy,lam),3);
% gradient
AFconcgrad = @(x,y,lam)=sum(Gfungrad(x,y,AFx,AFy,lam),3);

% (2) sample initial tipcells
% Model: all cells have lognormally distributed diameters, with mean =
% diam and std = sigma_diam, and one branch consists of NDdist such
% cells. The length of one such branch, i.e., the sum of NDdist
% i.i.d. such diameters, is approximated in the first two moments
% by a
% single lognormal random number.
\( r = \exp(\mu_{\text{branch}} + \sigma_{\text{branch}} \cdot \text{randn}(1)) \).
\( r(r \leq J_{\text{dist}}) = [\] ;
\( x = \text{cumsum}(r) ; \)
while \( x(\text{end}) < 1 \)
\( r = \exp(\mu_{\text{branch}} + \sigma_{\text{branch}} \cdot \text{randn}(1)) ; \)
if \( r > J_{\text{dist}} \), break; end
end
\( x = [x \ x(\text{end})+r] ; \)
end
\( x(x \geq 1-J_{\text{dist}}) = [\] ; % remove spill by end
\( x = [0 \ x]+\text{randn}(1) \cdot \text{rand} ; \) % place the origin \( \sim \text{U}(0,1) \)
\( x = [x(x \geq 1)-1 \ x(x < 1)] ; % \text{wrap} \ to \ (0,1) \)
\( y = \text{zeros}(\text{size}(x)) ; \) % bottom boundary
% (4) sampling of paths (Grow–Branch–Join)
% Model: tip cells can branch at an internal angle of theta_{\text{b}}, and
% otherwise grow in the direction of the gradient. Tip cells join if
% within J_{\text{dist}} distance of each other.

% number of tip cells and their index relative to previous layer
n_{\text{cells}} = numel(x) ;
ic_{\text{cells}} = 1-n_{\text{cells}}:0 ; % (counting backwards as it happens to be
convenient)
% the main algorithm output: vessel graph in between tip cells
A = \text{sparse}(n_{\text{cells}}, n_{\text{cells}}) ;
xy = [x(:, ) y(:, )] ;
% branching: none to start with
ib = \text{zeros}(1,0) ;
% vessel pairs exempt for joining in the next layer: none to start
% with (except that a vessel never joins itself)
exemptJ = \text{zeros}(2,0) ;
% "control points"/"ghost cells", collection of equispaced points
% at a
% per cell resolution on each vessel, used to find vessel crossings
XY = \text{complex}(x,y) ; % (complex numbers used since connect() is used)
ic = 1:n_{\text{cells}} ; % current list of indices of growing vessels
nextc = ic(\text{end}) ; % free index of cells starts at nextc+1
IC = ic ; % index of vessels in XY
% simulate until no active tip cells remain
layer = 1 ;
while n_{\text{cells}} > 0
% local physics
\( v = \text{AFconcgrad}(x,y,\lambda) ; \)
% duplicate tip cells at the branch points
[~,ibranch] = \text{fsetop}('ismember',ib,ic) ; % find them in ic
\( x = [x \ x(\text{ibranch})] ; \)
y = [y y(ibranch)];
n = numel(x);
icells = [icells icells(ibranch)]; % for connectivity to previous layer

inew = nextc+(1: numel(ibranch)); % new indices
ic = [ic inew]; % for connectivity to previous layer

nextc = ic(end);

% those tip cells selected for branching should not merge in this layer:
exemptJ = [exemptJ [ic(ibranch) ; inew] ;

if toggle_plot & visB
    plot(x(ibranch) , y(ibranch) , 'go' , 'MarkerFaceColor' , 'g');
end

% sample angles s.t. std(sigma_theta1−sigma_theta2) = sigma_thetab
sigma_theta1 = sigma_thetab*sqrt(6)*randgn(2).rand(size(ibranch)) ;
sigma_theta2 = sigma_thetab*sqrt(6)*randgn(2).rand(size(ibranch)) ;

% rotation of angles +/- thetab/2 relative to gradient:
v = [v v(ibranch).*exp(1i*(-thetab/2−sigma_theta2))];
v(ibranch) = v(ibranch).*exp(1i*(thetab/2+sigma_theta1));

v = v ./ av;

% proposed path update (xold,yold) ——> (x,y)

xold = x;
yold = y;
x = x + v0*real(v)+sigma0./av.*randgn(3).randn(size(x));
y = y + v0*imag(v)+sigma0./av.*randgn(3).randn(size(y));
y = min(y,1); % (cosmetic fix: upper boundary target was reached)

% add equispaced intermediate ”ghost” cells:
dXY = complex(xold , yold ) +(1: NDdist)' * complex(x−xold , y−yold ) /

% find all crossings requiring a merge by first finding all ghost cells that are close to each other:
[ii , jj] = find(connect(XY, Jdist)); % *** fsetop/unique faster

% at least one cell in the pair of close cells was just added (hence
% that particular vessel is growing) AND the two cells belong to
% *different* vessels:
id = find(jj ' > ixend & IC(ii) < IC(jj) | ...

% unique pairs of crossing vessels (note: the index of *the
% first*
% point of crossing is found by this construction):
[' ,iu] = fsetop('unique'.reshape(IC(id) ,2,[[]));
id = id(:,iu);

% *** can be avoided thanks to setdiff below

% remove vessels which are exempt for merging in this layer
[' ,iu] = fsetop('setdiff'.reshape(IC(id) ,2,[[]), exemptJ);
id = id(:,iu);
% merge into1 = merge id(2,j) into the farther grown vessel id(1, j),
% and vice versa for merge into2:
merge into1 = id(2,:) > ixend;
merge into2 = id(1,:) > ixend;
merge into1 = find(merge into1 & ~merge into2);
merge into2 = find(merge into2 & ~merge into1);

% merge vessels
z1 = XY(id(1,merge into1)+mod(ixend-id(1,merge into1),NDdist));
z2 = XY(id(2,merge into2)+mod(ixend-id(2,merge into2),NDdist));
% The above is essentially the same as
% z1 = XY(id(1,merge into1));
% but the construct picks the last point in the current vessel so as
% to stay at the main nodes in xy rather than at the ghost nodes in
% XY - this allows the graph to more easily be updated by the end
% (the arithmetics above hinges on the fact that mod(ixend,NDdist)
% is a constant).
% these vessels are discontinued; adjust their end positions here
[~, irem] = fssetop('ismember', IC(id(:,merge into1)), IC(id(:,merge into2)), ic);
x(irem) = real([z1 z2]);
y(irem) = imag([z1 z2]);
if visJMI
    plot(x(irem),y(irem), 'ks ');
end

% merge = two or more growing vessels merge into one point from
% which
% a single vessel continues to grow
merge = merge into1 & merge into2;
merge = find(merge);
% remove all crossings involving one vessel from ic(irem) as they
% will
% already have merged into another vessel and thus will be
% discontinued:
im = ~any(reshape(fssetop('ismember', reshape(IC(id(:,merge)), 1,[])), ic(irem)), 2, []);
merge = merge(im);

if isempty(merge)
    if toggle_plot && visJM
        plot(XY(id(:,merge)), 'bd-- ');
    end
end
% turning this on kills all merge (useful for debugging):
if false
    [~, im] = fssetop('ismember', reshape(IC(id(:,merge)), 1,[]), ic);
    irem_ = reshape(im, 2, []);
    irem = [irem irem_(:) '];
    icont = zeros(1,0);
    merge_cont = false(1,0);
else
    % find the connected components of the crossings
    [~, iu, ij] = fssetop('unique', reshape(IC(id(:,merge)), 1,[]));
    ij = reshape(ij, 2, []);
    % $$ L = sparse(ij(:,1), ij(:,2), 1, max(ij(:,1)), max(ij(:,2))) ; $$
end
% compute merge points as a connected component average
XY_ = reshape(XY(id(:,merge)),1,[]);
XY_ = XY_.*V;

% vessels that will continue
[ii, jj] = find(V); % ii = 1:numel(bins), jj = bins
[-, IU] = fsetop('ismember', ij(:,'), i'); % IU = ij?
jj = reshape(jj(IU),2,[]); % = bins(ij)
[-, IV] = fsetop('unique', jj(1, :)); % = 1:size(V,2)?
merge_cont = false(size(merge));
merge_cont(IV) = true;

% *** above can be seriously simplified if understood
% just merged vessels should not merge again in the next layer:
% exemptJ = reshape(IC(id(:,merge)),2,[]);
% exemptJ = [];
from_ = reshape(IC(id(:,merge)),1,[]);
for i = 1:numel(bins)
    bin = find(bins == i);
    n = size(bin, 2);
    for j = 1:sum(bin)
        k = j + 1:n
        pair = sort([from_(iu(bin(j))); from_(iu(bin(k))]);
        exemptJ = exemptJ + pair;
    end
end

% *** actually, the vessel which continues should not merge with
% anyone of the ones it just merged
if toggle_plot & & visJM
    plot(x(icont), y(icont), 'bd', 'MarkerFaceColor', 'b');
end
else
    % empty merge
    icont = zeros(1, 0);
    merge_cont = false(1, 0);
    exemptJ = zeros(2, 0);
end

% regenerate ghost cells since end-points have changed
\[ dXY = \text{complex}(x_{\text{old}}, y_{\text{old}}) + (1:NDdist) \ast \text{complex}(x-x_{\text{old}}, y-y_{\text{old}}) / NDdist; \]

\[ XY(ixend+1:end) = \text{reshape}(dXY, [1]); \]

\% *** only regenerate those affected!

\% the corresponding update to the graph
\[ xy = [xy; \text{x}(:, y(:))]; \]
\[ A = [A; \text{sparse}(1:n_{\text{cell}}, i_{\text{cell}}+size(A,2), 1, n_{\text{cell}}, size(A,2))]; \]

\% vessels to discontinue:
\[ irem = \text{fsetop('union', irem, \text{reshape}(\text{find}(y >= 1), 1, []))}; \]
\[ x(irem) = []; \]
\[ y(irem) = []; \]

\% sample vessels for branching in the next layer:
\[ ib = ic; \]
\[ ib([irem icont]) = []; \% (can not branch: removed vessels or just joined) \]

if RAND
\[ ib(\text{randgn(4), rand(size(ib)) > Pbranch}) = []; \]
else
\% without randomness this is an attempt to visually spread the
\% branching evenly according to the parameter Pbranch:
\[ ib = ib(\text{diff(mod(cumsum([layer Pbranch repmat(Pbranch, size(ib)) \])), 1)) < 0); \]
\[ ib = \text{reshape}(ib, 1, []); \]
end

\% remaining vessels
\[ ic(irem) = []; \]
\[ icell = 1-n_{\text{cell}}:0; \]
\[ icell(irem) = []; \% (index of current cells relative to previous layer) \]
\[ n_{\text{cell}} = \text{numel(ic)}; \]
\[ layer = layer+1; \]
end

\% there are double nodes in the graph which we now join:
\[ [\text{~}, iu, id] = \text{fsetop('unique', round(xy*1e6))}; \]
\[ xy = xy(iu,:); \]
\[ [ii, jj, ss] = \text{find}(A); \]
\[ A = \text{fsparse}(id(ii), id(jj), ss', size(A)); \]

\[ [ii, jj] = \text{find}(A); \]
if toggle_plot
\[ \text{gplot}(A, xy, 'r-'); \]
\[ \text{plot(digraph(jj, ii), 'r-'), 'XData', xy(:,1), 'YData', xy(:,2)); \]
\[ \text{plot(AFx(:,), AFy(:,), 'k*'); \]
\[ \text{plot(XY, 'k-'); \]
\[ \text{axis([0 1 0 1]);} \]
\[ \text{axis square} \]
end

if toggle_plot && visAF
\% add the AF gradient field
\[ x0 = \text{linspace}(0,1,11); x0([1 \text{ end}]) = []; \]
\[ y0 = x0'; y0(\text{end}) = []; \]
\[ [X0, Y0] = \text{meshgrid}(x0, y0); \]
\[ G = \text{AFconcgrad}(X0, Y0, lam); \]
\[ \text{quiver}(X0, Y0, \text{real}(G), \text{imag}(G), 2, 'color', [0.05 0.05 0.75]); \]
pause
end