A Branching Process Approach To The Identification of Influential Nodes in Complex Networks

by

Ailbhe Cassidy

Under the supervision of

Prof. James P. Gleeson and Dr. Ali Faqeeh


August 2019
Abstract

We live in a world surrounded by networks. They are ubiquitous. Be it social media networks linking us to our friends, transportation networks interconnecting locations around the globe, or the metabolic network breaking down food in our bodies. The study of networks is an interdisciplinary field spanning from fields of mathematics, psychology, sociology, biology, computer science, physics, and many other areas. The field of Network Science has greatly profited from the contributions of such diverse scientific communities. However, there are still remaining challenges that are open for further research and discussion. The identification of influential nodes in a network is a constant challenge faced by researchers. Regardless of the specific field of study the solution to this problem is constantly in demand.

In this thesis, we present a new measure for the identification of influential nodes in complex networks. It is based on a mathematical model which uses a branching process approach. Unlike a lot of existing measures, it is based on a mathematical model that takes into account not only the structure of the network but also the dynamics taking place on the network. We present the mathematical theory behind the model and explain from this where the measure will return accurate results, and when it should return inaccurate predications. Throughout this work, we provide a considerable amount of results on a range of networks. We do this to support our proposal and recommendations for the usage of this new centrality metric.
Acknowledgements

I would like to take this opportunity to thank my supervisor Prof. James Gleeson, for his continuous advice and patience during the preparation of this work. I graciously acknowledge the help and guidance of Dr. Ali Faqeeh. Thank you very much for your words of encouragement during the year. I could not have asked for a better mentor. I would also like to thank my colleagues and friends in MACSI, especially my office buddies in A2016g.

Additionally, I would like to thank my family for their ongoing support. I acknowledge my dad for initially instilling a love for maths in me, and his help during my school years. I especially warmly thank my mam Elaine. You put everything into your children and without you this would not be possible.

This work was funded by Science Foundation Ireland (grant no.16/IA/4470).
## Contents

1 Introduction .......................................................... 1
   1.1 General Introduction ........................................... 1
   1.2 Network Models ................................................. 2
     1.2.1 Networks And Their Representation ....................... 3
     1.2.2 Random Graphs ............................................ 5
     1.2.3 The Configuration Model .................................. 8
   1.3 Epidemic Models on Networks ................................. 9
     1.3.1 The SI Model ............................................. 10
       1.3.1.1 The 'Fully Mixed' SI Model .......................... 10
       1.3.1.2 The SI Model on a Network .......................... 12
     1.3.2 The SIR Model ............................................ 13
       1.3.2.1 The 'Fully Mixed' SIR Model .......................... 13
       1.3.2.2 The SIR Model on a Network .......................... 16
     1.3.3 The SIS Model ............................................ 19
       1.3.3.1 The 'Fully Mixed' SIS Model .......................... 19
       1.3.3.2 The SIS Model on a Network .......................... 20
   1.4 Implementing Disease Spreading Models ..................... 20
     1.4.1 The Independent Cascade Model .......................... 21
     1.4.2 The Gillespie Algorithm ................................ 22
     1.4.3 Discrete-Time Versus Continuous-Time ................... 23

2 Centrality .......................................................... 27
   2.1 What is Centrality? ............................................ 27
   2.2 Centrality Measures .......................................... 27
     2.2.1 Degree Centrality ........................................ 27
     2.2.2 Eigenvector Centrality .................................. 28
     2.2.3 Katz Centrality ........................................... 29
     2.2.4 PageRank Centrality ..................................... 29
     2.2.5 Closeness Centrality .................................... 30
     2.2.6 Betweenness Centrality .................................. 30
     2.2.7 K-Shell Decomposition ................................... 31
   2.3 The Non-Backtracking Centrality ............................. 31
     2.3.1 The Non-Backtracking Matrix ............................. 32
     2.3.2 The Non-Backtracking Centrality ......................... 34
   2.4 A Review of the Literature: Identification of Influential Nodes ... 35

3 A New Centrality Measure: The Branching Process Centrality .......... 43
   3.1 Classical Branching Processes ............................... 43
     3.1.1 Probability Generating Functions ......................... 43
     3.1.2 Galton-Watson Branching Processes ....................... 44
   3.2 A Mathematical Model: Predicting the Most Influential Nodes in a Network ......................................................... 45
   3.3 Implementation of the Branching Process Centrality Measure ........ 48
   3.4 Branching Process Centrality Versus Other Centrality Measures ... 50
Chapter 1

Introduction

1.1 General Introduction

The role and contribution of networks to our understanding of real-world systems has consequently attracted the attention of researchers in various fields. Networks can describe various types of interactions, spanning from individuals meeting, the spreading of rumours or ideas amongst people or even the propagation of diseases within a community [1]. The study of networks allows us to explain such interactions. Networks are a powerful tool that allow for the comprehension of complex systems [2]. For that reason, the entire research area of Network Science is solely dedicated to the study of networks.

Generally, the roles played by individual entities (nodes) in a complex network can be quite diverse. The ability to understand the impact of individual nodes on the global functionality of the system is currently a fundamental problem in Network Science.

If we think of any real-world network, it would seem obvious that many mechanisms on the network could be highly affected or controlled by a fraction of influential nodes. The identification of influential nodes that lead to faster and wider spreading in complex networks is of great significance.

In the case of news spreading through a network, locating these ‘super-spreaders’ and targeting them could potentially help in optimizing the availability of resources and ensuring the most affective spread of information. Conversely, in order to prevent the vast spreading of a disease on a network, these influential spreaders are good choices for efficient vaccination with hopes to reduce the overall size of an epidemic. Targeting or secluding the influential nodes can be advantageous when attempting to control the overall spreading process taking place on a network.

The work of this thesis is focused on the identification of the influential nodes in a network. To date, there exists many mathematical measures which are available for the classification of nodes in terms of their importance. In Network Science, these are known as centrality measures and each measure assigns a centrality value (a number indicating importance) to each node in the network. In the context of spreading processes, the nodes assigned the highest numbers by the centrality measure are classified as the ‘super-spreaders’ in the network.

There does not exist a perfect centrality measure. Particular measures are more accurate in certain situations. This might depend on the system that is being represented by the network, the structure and topology of the network, ones definition of importance etc. There is not one global measure that we can use for the identification of efficient spreaders on a network. Many of the measures that exist have been proposed through intuitive and physical reasoning. In this thesis, we propose a new centrality
measure based on a mathematical model. We suggest situations where we believe it produces dependable and accurate predictions and also, situations when it is likely to fail.

To begin the thesis, we introduce some basic concepts of Network Science. In this chapter, we start by discussing some of the important notation and mathematical concepts in the area. This is a broad subject and so, we only deal with key topics that we believe are necessary for the successful comprehension of the material in this thesis. We discuss some of the popular models for network structure. We explain their role and the reasons many scientists often lean on synthetic networks first, when posed with a research question. We then, move on and discuss some of the models that exist for the spreading dynamics that take place on networks. To finish this chapter, we explain how we implement the spreading models and the difference between taking a discrete-time and continuous-time approach when simulating the processes.

In Chapter 2, we explore in detail the various centrality measures that can be found in the literature. We give the mathematical description of each centrality measure and point out several downfalls associated with the different measures. We then, discuss various results given in the literature surrounding the area of ‘super-spreader’ identification. Here, we note the agreements and disparities amongst the interpretation of results in the various papers.

Once we go through the necessary background material and define the currently available centrality measures, we present our newly proposed centrality measure. In Chapter 3, we define our new centrality measure mathematically and present some results. In Chapter 4, we consider the performance of the new centrality measure on social networks with dynamics that are characteristic of these networks. In these chapters, we justify the new centrality measure’s possible superiority in certain situations over other centrality measures by revealing various results.

1.2 Network Models

A network is a collection of discrete objects joined together by links. In Network Science, these objects are referred to as nodes and the links are referred to as edges [1]. Many systems in everyday life, be it social, biological, or technological can be viewed as networks. In fact, thinking of these systems as networks can lead to useful insights into the particular area of interest.

We encounter networks of one kind or another all of the time [3]. A network can represent the connections between people on Instagram or Facebook, the connections between neurons in the brain, the connections between different web pages on the World Wide Web. The list goes on. Networks are everywhere, and this is a reason why a lot of research is being undertaken in the field of Network Science.

Regardless of the specific system being studied, the same communal question arises. If we know the structure of a system’s network, what can we learn about the system? How are the structural features of a network related to the qualities of the system we are
Networks represent the pattern of interactions between objects in a system. This pattern of connections has an impact on the behaviour of that system. For example, the pattern of connections in a social network affects the spread of information, gossip, rumours. Unless we know the underlying connections in the network, we will not understand how the overall system behaves. One of the best ways to learn about the effects of different structures on the overall system’s behaviour is to build mathematical models.

In the next few sections, we will introduce some of the most widely used models of network structure, closely following the description in [1]. These models are used to mimic some of the patterns of connections in real-world networks. They allow the creation of artificial networks with known parameters. This gives insight into the structure of networks and how changing parameter values affects the shape of the network. These artificial models also answer our questions by providing an understanding of processes taking place on networks with certain structures. The investigation of properties and behaviours of such synthetic networks helps the understanding of similarly structured real-world networks.

1.2.1 Networks And Their Representation

Before introducing some of the fundamental network models, we present some concepts and mathematical background that is needed when studying networks [4] [5] [6]. As we have said, a network is a collection of nodes joined together by edges. The adjacency matrix of a network includes the information on the connections between nodes. It is the fundamental mathematical representation of a network. The adjacency matrix $A$ of an undirected network with $n$ nodes, is defined to be the $n \times n$ matrix with elements $A_{ij}$ such that

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases} \quad (1.1)$$

![Figure 1.1: An example of an undirected network.](image)
For example the adjacency matrix for the graph in Figure 1.1 is given by

$$A = \begin{pmatrix}
0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 & 0
\end{pmatrix}. \quad (1.2)$$

For directed networks, each link has a direction from one node to another but the link is not necessarily reciprocated in the opposite direction. The elements of the adjacency matrix for a directed network are given by

$$A_{ij} = \begin{cases}
1 & \text{if there is a directed edge from node } i \text{ to } j, \\
0 & \text{otherwise.} 
\end{cases} \quad (1.3)$$

In directed networks, the adjacency matrix is not symmetric. However, for undirected networks the adjacency matrix will always be symmetric. This is because the edge from node $i$ to node $j$ is the same as the edge from node $j$ to node $i$.

![Directed Network Diagram](image)

Figure 1.2: An example of a directed network.

The adjacency matrix for the directed network in Fig 1.2 is given by

$$A = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0
\end{pmatrix}. \quad (1.4)$$
Another important property is the degree of the nodes in a network. The degree of a node in an undirected network is the number of edges connected to it. For an undirected network of \( n \) nodes, the degree of a node \( i \) can be written as
\[
k_i = \sum_{j=1}^{n} A_{ij}.
\]  

A related quantity is the mean degree of a node \( \langle k \rangle \) in an undirected network. This is given by
\[
\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i.
\]  

If there are \( E \) edges in a network then there are \( 2E \) ends of edges (stubs) in the network. We have that \( \sum_i k_i = 2E \), and the average degree of a node can be written as
\[
\langle k \rangle = \frac{2E}{n}.
\]  

Lastly, we will finish this section with the definition of a specific network structure that will be important for future work in this thesis. A tree is a connected, undirected network that contains no loops [7] [8] [9]. By connected, we mean that every node in the network is reachable from every other via some path in the network.

![Example of a tree structured network.](image)

Since trees have no closed loops, there is exactly one path between any pair of nodes in the network. This property makes certain kinds of calculations simple, and for that reason, trees are sometimes used as basic network models. Later in this thesis we will use the term ‘locally’ tree-like to describe some networks. This means that in certain regions of the network the structure is that of a tree.

### 1.2.2 Random Graphs
A random graph is a network model in which the values of certain properties are fixed but the network is otherwise completely random [10]. The Erdös-Rényi model is either
of two closely related models for generating random graphs [11] [12] [13] [14]. In the deterministic model of Erdős-Rényi, we fix the number of nodes \( n \) and the number of edges \( E \) in the network. Then, the \( E \) edges are placed among the \( n \) nodes at random.

Alternatively, the network can be formed stochastically (randomly) where we fix not the number but the probability of edges between nodes. As before, we have \( n \) nodes in our network but now we place an edge between every pair of nodes with probability \( p \).

The expected number of edges in the stochastic Erdős-Rényi graph \( G(n, p) \) is given by \( \binom{n}{2} p \). Therefore, in order to have the same (expected) number of edges in the deterministic Erdős-Rényi graph \( G(n, E) \) and the stochastic \( G(n, p) \) Erdős-Rényi graph, we must have

\[
\binom{n}{2} p = E, \\
\Rightarrow p = \frac{2E}{n(n-1)}.
\]

The probability of two random nodes being connected in \( G(n, p) \) must satisfy equation (1.9) in order to coincide with the deterministic Erdős-Rényi model \( G(n, E) \).

As previously seen in equation (1.7), the mean degree of a node in a network with precisely \( E \) edges is \( \frac{2E}{n} \). So, the mean degree \( \langle k \rangle \) in the stochastic random graph with \( n \) nodes is given by

\[
\langle k \rangle = \left( \frac{2E}{n} \right) = \frac{2\langle E \rangle}{n} = \frac{2}{n} \binom{n}{2} p = \frac{2}{n} \cdot \frac{n(n-1)}{2} p = (n-1)p. 
\]

The degree distribution can also easily be calculated for the stochastic Erdős-Rényi graph. Each node in the network is connected to the other \( n-1 \) nodes with a probability \( p \). The probability of being connected to \( k \) other nodes but not the others is \( p^k(1-p)^{n-1-k} \). There are \( \binom{n-1}{k} \) ways to choose these \( k \) other nodes, thus the probability of being connected to exactly \( k \) nodes is given by

\[
p_k = \left( \frac{n-1}{k} \right) p^k (1-p)^{n-1-k}. 
\]

This is a binomial distribution. Therefore, the Erdős-Rényi random graph has a binomial degree distribution.

We are generally interested in large real-world networks, and these networks tend to have a large number of nodes \( n \). Most real-world networks are also sparse. Their mean degree \( \langle k \rangle \) increases slower than \( n \) as \( n \) becomes large. This implies that \( p = \frac{\langle k \rangle}{n-1} \) becomes very small. By taking the natural logarithm of \( (1-p)^{n-1-k} \), we can write

\[
\ln[(1-p)^{n-1-k}] = (n-1-k)\ln\left(1 - \frac{\langle k \rangle}{n-1} \right). 
\]
By expanding the natural logarithm as a Taylor series and letting $n \to \infty$, we get

$$\log\left(1 - \frac{\langle k \rangle}{n-1}\right) = -(n - 1 - k)\frac{\langle k \rangle}{n-1} = -\langle k \rangle.$$ \hfill (1.13)

Now taking the exponential, we get

$$(1 - p)^{n-1-k} = e^{-\langle k \rangle}.$$ \hfill (1.14)

For large $n$, we also have

$$\left(\begin{array}{c} n-1 \\ k \end{array}\right) = \frac{(n-1)(n-2)\cdots(n-k)}{k!} \approx \frac{(n-1)^k}{k!}.$$ \hfill (1.15)

Thus, the degree distribution in the limit of large $n$ can be written as

$$p_k = \frac{(n-1)^k}{k!} p^k e^{-\langle k \rangle} = \frac{(n-1)^k}{k!} \left(\frac{\langle k \rangle}{n-1}\right)^k e^{-\langle k \rangle} = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}. \hfill (1.16)$$

This is a Poisson distribution. So, in the limit of large $n$, the stochastic Erdős-Rényi random graph has a Poisson degree distribution.

The Erdős-Rényi random graph is both simple to describe and straightforward to study using analytic methods, as we have demonstrated above. For this reason, it would seem to be an excellent model for investigating all sorts of network phenomena. However, the model has some major drawbacks. It is unlike real-world networks in various ways.

The cluster coefficient $C$ of a network is commonly defined as the probability that two network neighbours of a node are also neighbours (presence of short loops such as triangles in the network) [15]. In real-world networks, there is often considerably large clustering coefficients [1]. In the Erdős-Rényi random graph $G(n, p)$ (equation (1.10)), the probability that any two nodes are neighbours is the same $p = \frac{\langle k \rangle}{n-1}$. Hence,

$$C = \frac{\langle k \rangle}{n-1}. \hfill (1.17)$$

In the limit $n \to \infty$ in sparse networks this tends to zero. This is an immediately obvious shortcoming of the random graph model.

Another considerable property where real-world and random networks diverge is their degree distribution. Real-world networks tend to have a right-skewed degree distribution [1], that is most nodes have low degree and a small number of nodes have very high degrees. The random graph, in the limit of large $n$, has a Poisson degree distribution.
which is not right-skewed.

These contrasting properties have an impact on the structure of the network. The differing patterns of connections in these networks will result in different behaviour. Thus, possibly making the random graph an inaccurate model for processes taking place on real-world networks.

1.2.3 The Configuration Model

The configuration model addresses the shortcomings of the Erdös-Rényi random graph with respect to the degree distribution. Instead of being limited to a Poisson degree distribution, the configuration model can be specified to have any degree distribution.

The configuration model is a random graph which takes a degree sequence as its input. This degree sequence is derived from the specified degree distribution. Thus, the degree of every node is given. A random network can be created using these degrees. Each node $i$ has the same number of stubs as its degree $k_i$. Two stubs in the network are chosen at random and connected together to form an edge. This process is repeated until no unconnected stubs remain.

Using this process, the network may contain self-edges [16] or multi-edges [17]. These are not characteristic of many real-world networks. However, in the limit of large network size, the number of self-edges and multi-edges becomes a negligible fraction of the overall network [1], hence not having much of an overall effect.

Many real-world networks are claimed to have a power-law degree distribution [18], (there are recent arguments that refute this finding [19], but we don’t get into it here!). In these networks, there are many nodes with low degrees and only a few nodes have very high degrees in the tail of the distribution [20]. We could begin to understand some of the important properties of real-world networks by studying the configuration model with a power-law degree distribution.

As we did for the Erdös-Rényi random graph, we calculate the clustering coefficient for the configuration model. We denote $p_k$ to be the fraction of nodes in the network with degree $k$. If there are $n$ nodes in the network, then there are $np_k$ nodes with a degree of $k$. We also know from equation (1.7) above, that there are $\sum_i k_i = 2E$ stubs in the network. Since each node has an equal chance of being attached to a free edge of every other node then, the probability that a randomly chosen node in the network is attached to a degree $k$ node by one of its edges is therefore $\frac{k}{2E-1}$. If there are a lot of edges in the network, we can write $\frac{k}{2E-1} \approx \frac{k}{2E}$. Thus, the probability of our edge attaching to any node with degree $k$ is

$$\frac{k}{2E} \times np_k = \frac{k np_k}{\langle k \rangle}.$$  (1.18)

Now, we want to calculate the clustering coefficient which is the probability that two network neighbours of a node are also neighbours. Consider a node $l$ with at least two
neighbours. We call two specific neighbours of \( l, i \) and \( j \). Both \( i \) and \( j \) are at the other ends of edges from \( l \). Excluding these edges, the number of other edges connected to them are denoted by \( k_i \) ans \( k_j \). These are the excess degrees [21] of nodes \( i \) and \( j \). The probability \( q_k \) of having an excess degree \( k \) is equal to the probability of having a total degree of \( k + 1 \), and so using equation (1.18), we get

\[
q_k = \frac{(k + 1)p_{k+1}}{\langle k \rangle}.
\] (1.19)

The probability of there being an edge between \( i \) and \( j \) is just \( k_i k_j / 2E \). We average both \( k_i \) and \( k_j \) over the distribution \( q_k \) to get the clustering coefficient

\[
C = \sum_{k_i, k_j = 0}^{\infty} q_{k_i} q_{k_j} \frac{k_i k_j}{2E} = \frac{1}{2E} \left[ \sum_{k=0}^{\infty} k q_k \right]^2 = \frac{1}{2E \langle k \rangle^2} \left[ \sum_{k=0}^{\infty} k (k + 1) p_{k+1} \right]^2 = \frac{1}{2E \langle k \rangle^2} \left[ \sum_{k=0}^{\infty} (k - 1) k p_k \right]^2 \neq \frac{1}{n} \frac{(\langle k^2 \rangle - \langle k \rangle)^2}{\langle k \rangle^3}.
\] (1.20)

Therefore the clustering coefficient is given by

\[
C = \frac{\langle k \rangle}{n} \left[ \frac{(\langle k^2 \rangle - \langle k \rangle)}{\langle k \rangle^2} \right]^2.
\] (1.21)

Thus for the configuration model, the clustering coefficient \( C \) vanishes as \( n \to \infty \). However, for finite \( n \) with large \( \frac{(\langle k^2 \rangle - \langle k \rangle)}{\langle k \rangle^2} \), which is the case for networks with power law distribution, we can arrive at large values for the clustering coefficient.

The configuration model is an important tool when studying networks. It’s a generally simple model to deal with, while also possessing features of real-world networks. For this reason, it is often the model researchers go to first when studying a new process or question.

### 1.3 Epidemic Models on Networks

In recent years, there has been a large increase in the investment made by the scientific community into the study of social networks [22] [23] [24]. A reason for this, is the connection between social networks and the spread of diseases [25]. The connection is simple. Diseases spread over networks of contacts between people. For example, some diseases such as the Influenza or the Measles are transmitted when two people breathe the same air. Other diseases are transmitted when people touch and sexually transmitted diseases are transferred from one person to the other when they have sex [26] [27]. The
transmission of any disease from an individual to another, is solely dependent on that
individual's contacts. The pattern of such contacts can be represented as networks. An
individual with the flu may communicate the disease to a work colleague that shares the
same office as opposed to a friend living in a different county [28] [29]. Knowing these
social network structures allows us, along with spreading dynamics of the disease, to
understand and predict the outcomes of epidemics [30].

There are other spreading processes on networks that can be compared to the spreading
of diseases. The spreading of information, news and gossip through a community has
properties similar to the properties associated with the spreading of a disease through
a population. An individual may hear a piece of information or news from their neigh-
bour, that individual now being ‘infected’ can pass the news onto their neighbours.
The adoption of ideas or trends for example may also be contagious in this way. Many
researchers in the literature in fact, use the ideas and models for the spread of disease
and apply them to help the understanding of information spread.

In order to represent an epidemic mathematically, the dynamics of the disease are sim-
plified to just changes between a few disease states. Mathematical models incorporating
these states are called compartment models where the individual states are called com-
partments [31] [32].

A mathematical modelling approach to epidemics predates the study of networks by
many years [33] [34]. The first work in the area was performed by Anderson McK-
endrick, a medical doctor and amateur mathematician. The traditional approach didn’t
consider network structures. It made use of the ‘fully mixed’ approach. This approach
to epidemic modelling assumes that every individual has an equal chance, per unit time,
of coming in contact with every other. It assumes people come in contact completely at
random. This is of course not realistic.

In reality, people do not have contact with the whole population; in fact, an individual
would have contact with only a small portion of the population, which isn’t random. An
individual will have contacts with their family, friends, work colleagues, etc. Taking
the population of the world into consideration, this individuals contacts are only a tiny
fraction of the whole population. For this reason, networks play an important role in
the spread of disease and other things [35]. However, the traditional approaches are still
useful in our study of spreading dynamics on networks. In the following section, we
will introduce three different disease spread models from the traditional ‘fully mixed’
approach. We will then, show that including the structure of the network has a strong
effect on the disease spreading process.

1.3.1 The SI Model

1.3.1.1 The ‘Fully Mixed’ SI Model

In this section, we will focus on a simple compartmental model where there are just two
states. The SI model only considers two states, susceptible and infected. A person is
considered to be in the susceptible state in the model if they do not have the disease but
they have the potential of contracting it, if they come in contact with someone with the
disease. A person in the infected state is someone that has the disease and that could possibly pass it onto someone in the susceptible state if they become in contact. This model is a gross simplification of what happens in reality, however it captures various important features of disease dynamics and it useful for understanding what is happening on the network level.

Consider the ‘fully mixed’ approach, with a disease spreading through a population of individuals. We let $s(t)$ and $i(t)$ be the average fraction of susceptible and infected individuals in the network at time $t$. We note that the spreading process is a random one. For that reason, we use the average fraction of nodes over many realisations of the process. It is obvious that the fraction of infected individuals goes up as the susceptible individuals contract the disease. We take $\beta$ to be the random per-individual rate at which individuals in the network make a contact that leads to capturing the disease. Therefore, an infected person has such a contact with an average of $\beta s$ susceptible people per unit time. Since there is on average a fraction $i$ of individuals that are infected, then the overall average rate of new infections will be $\beta i s$, and we can write a differential equation for the rate of change of $i$ as

$$\frac{di}{dt} = \beta si,$$  \hspace{1cm} (1.22)

with $s + i = 1$, we have

$$\frac{di}{dt} = \beta(1 - i)i.$$  \hspace{1cm} (1.23)

This simple mathematical model for the spread of a disease is called the fully mixed SI model. The solution to this equation is given by

$$i(t) = \frac{i_0 e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}},$$  \hspace{1cm} (1.24)

where $i_0$ is the value of $i(t)$ at $t = 0$ [1]. The solution (1.24) to the differential equation (1.23), produces an S-shaped curve for the average fraction of infected individuals. At short times, the curve increases exponentially corresponding to the initial phase of the disease in which most people are susceptible. The curve then saturates as the fraction of susceptible individuals decreases, as there are less and less victims remained to be infected.
1.3.1.2 The SI Model on a Network

The traditional approach to epidemic modelling assumes that any individual can have contact and potentially transmit a disease to any other. This is not an accurate assumption. In reality, people have a set of regular acquaintances whom they have contact with at some frequency and the rest of the population can be ignored. The patterns of these contacts can be represented by networks, and the structure of these networks has a strong impact on the way the disease spreads.

In contrast to the ‘fully mixed’ approach, network models make use of the network of contacts between individuals. In these models, the transmission rate $\beta$ is no longer the rate of contacts between an infected individual and all others in the population. It is now the rate in which infection will be transmitted between two individuals, one infected and one susceptible, who are connected by an edge in the network.

In the network version of the SI model, the individuals are represented by nodes. Infected nodes spread the disease to their susceptible neighbours at a rate $\beta$. In this model, infected nodes remain infected forever so over time the disease spreads throughout the network.

Generally, we simulate the model on a computer to observe the dynamics of how the disease spreads. However, we can simply calculate the total size of the disease outbreak. In the limit of long time, every individual that can be infected by the disease is infected. If there is a path between the initially infected individual and another, that node will eventually become infected too.

It is well known that a lot of networks are composed of one large component and a few smaller components [1]. If we start with one initially infected node in the large component, in the long time limit, the disease will spread throughout the entire big component. This leads to a large outbreak of the disease. However, if the initially infected node is
in one of the small components, the resulting behaviour is only a small outbreak of the disease.

We can now see an obvious difference between the ‘fully mixed’ model and network models. In the ‘fully mixed’ model, there were also two possibilities - a large outbreak of a disease or a small one that eventually dies out. The outcome was determined by the model and its parameters. In the network model, there is a new stochastic element involved. The outcome is now determined by the model and its parameters but also dependent on the network structure and the position of the initially infected node in the network. Network models introduce a randomness, where if we repeat the process using the same model parameters, we might get either outcome. By looking at the SI model, we are given an insight into the impact of the networks structure on the disease spreading process.

1.3.2 The SIR Model

1.3.2.1 The ‘Fully Mixed’ SIR Model

One way to make the SI model more realistic is to add a recovery compartment. This allows an individual that is infected with a disease to recover. This extension to the SI model is reasonable because it addresses the case where an individuals immune system might fight a disease and subsequently recovers. Also for some particular diseases people often retain their immunity to it after they recover, meaning they cannot catch it again. This extra recovery state allows the modelling of such scenarios. This model with three states is called the SIR model.

Firstly, we consider the ‘fully mixed’ approach to the SIR model. The dynamics of the model are as follows. Initially, a susceptible individual becomes infected when it has infectious contact with an infected individual. As before, such contacts happen at an average rate $\beta$ per person. An infected person then can recover at some constant rate $\gamma$.

The average fractions $s$, $i$ and $r$ of individuals in these states is given by

$$\frac{ds}{dt} = -\beta si, \quad (1.25)$$

$$\frac{di}{dt} = \beta si - \gamma i, \quad (1.26)$$

$$\frac{dr}{dt} = \gamma i, \quad (1.27)$$

where $s + i + r = 1 \ [1]$. In practice we cannot solve this in closed form. However, it can be evaluated numerically in order to plot the three curves for $s$, $i$ and $r$ versus time.
The fraction of susceptible individuals decreases with time as susceptible people become infected. The fraction of infected increases at first as susceptible people get infected but then decreases to zero in the long time limit because they recover. The fraction of susceptible individuals may not go to zero. This is because when the curve for $i(t)$ goes to zero, there are no more infected individuals left to infect the possibly remaining susceptible people. Also in this case, the fraction of recovered people would not quite reach one in the long time limit.

![Figure 1.5: The time evolution of the SIR model. Figure based on plots in [1].](image)

The asymptotic behaviour of $r$ is important. It is the total fraction of individuals that caught the disease during the epidemic. This tells us the total size of the epidemic. This can be calculated directly by finding the steady state of equation (1.27)

$$\frac{dr}{dt} = 0.$$  \hspace{1cm} (1.28)

Therefore

$$\gamma i = 0,$$  \hspace{1cm} (1.29)

where $i = 1 - s - r$. From equation (1.25)

$$\frac{ds}{dt} = -\frac{\beta s}{\gamma} \frac{dr}{dt}.$$  \hspace{1cm} (1.30)

Rearranging to get
\[
\frac{1}{s} \frac{ds}{dt} = -\frac{\beta}{\gamma} \frac{dr}{dt}.
\] (1.31)

Using the value \(s_0\) as the value of \(s(t)\) at \(t = 0\), and taking into account that nobody has recovered at \(t=0\), we integrate both sides to get

\[
s = s_0 e^{-\frac{\beta r}{\gamma}}.
\] (1.32)

We therefore have

\[
\gamma (1 - r - s_0 e^{-\frac{\beta r}{\gamma}}) = 0.
\] (1.33)

We can rearrange this to get

\[
r = 1 - s_0 e^{-\frac{\beta r}{\gamma}}.
\] (1.34)

This is the equation for the total size of the outbreak. Assuming the disease initially starts with one infected individual, we can write

\[
r = 1 - e^{-\frac{\beta r}{\gamma}}.
\] (1.35)

This cannot be solved in closed form but we can plot \(r\) versus \(\frac{\beta}{\gamma}\) to see what the size of the outbreak looks like as a function of the parameters \(\beta\) and \(\gamma\) (solved numerically by making an initial guess and iterating the equation to convergence [1]).

Figure 1.6: An example curve for the fraction of nodes recovered as a function of \(\frac{\beta}{\gamma}\). Figure based on plots in [1].
Figure 1.6 above is an example of such a curve. Depending on the system, this curve will vary slightly. In general for $\frac{\beta}{\gamma} \leq 1$, we have $r$ close to zero and have no epidemic at all. This makes intuitive sense. The infected individuals recover faster than the susceptible individuals get infected, and so the disease dies out. At the point $\beta=\gamma$, the size of the epidemic suddenly increases from zero. This transition point is called the ‘epidemic threshold’ or the ‘critical point’. This is an interesting property which we have derived solely using the ‘fully mixed’ approach.

Another quantity of interest could be the length of time an individual remains infected in the system. For the SIR model, if we know the recovery rate we can calculate the length of time $\tau$, that an infected individual is likely to remain infected before recovery.

The probability of recovering in a time interval $dt$ is $\gamma dt$, and therefore the probability of not recovering is $1 - \gamma dt$. Thus, the probability of still being infected after a total time $\tau$ is

$$\lim_{dt \to 0} (1 - \gamma dt)^{\tau} = e^{-\gamma \tau}.$$ (1.36)

The probability $P(\tau)$ that the individual remains infected for time $\tau$ and then recovers in the interval $(\tau, \tau + dt)$ is

$$P(\tau) = \gamma e^{-\gamma \tau} dt.$$ (1.37)

This is an exponential distribution. It implies that an infected individual is most likely to recover immediately after becoming infected but may remain infected for a long time. This is not ideal. With real diseases, individuals normally remain infected for about the same amount of time and only a few people recover much earlier or much later than this average time. Here, we can see an immediately obvious issue that arises from taking the ‘fully mixed’ approach to the SIR model.

1.3.2.2 The SIR Model on a Network

We will now consider the SIR model on a network. In this model individuals can recover so unlike the SI model, it is no longer true that a susceptible neighbour of an infected node will eventually become infected.

The probability of infection in the time interval $(t, t + \tau)$ is given by

$$1 - \text{prob(not infected in the time interval} (t, t + \tau)),

= 1 - \prod_{i=1}^{m} (1 - \beta dt),

= 1 - (1 - \beta dt)^m,$
\[ = 1 - \left(1 - \frac{\beta \tau m}{m}\right)^m, \quad (1.38) \]

where \( m = \frac{\tau}{m} \). Taking the limit as \( m \to \infty \), we have \( \lim_{m \to \infty} \left(1 - \frac{\beta \tau m}{m}\right)^m = e^{-\beta \tau} \). Thus the transmission probability is given by

\[ \phi = 1 - e^{-\beta \tau}. \quad (1.39) \]

Here we generally assume \( \tau = \text{constant} \), that is every infected individual remains infected for the same length of time. As discussed before, this is more realistic than the exponential distribution of \( \tau \) that arose from considering the ‘fully mixed’ model. So now, the transmission probability is constant throughout the network. The SI model on a network can be considered a particular case of this, with \( \phi = 1 \).

As discussed in the previous section, we observe the existence of an epidemic threshold solely by considering the ‘fully mixed’ approach to the SIR model. Now, we calculate the position of the epidemic threshold in terms of the transmission probability \( \phi \) on a network for the SIR model. We do this by using a branching process approach. An important point to note is that, in this calculation we assume the children nodes are independent and therefore we are assuming the network is structurally tree-like.

Figure 1.7: Initially infected node (black) connected by an edge to a susceptible node (white).

Figure 1.8: Initially infected node infects its neighbour. There are now four vulnerable children to the disease.

We begin with one initially infected node in a network (Figure 1.7), with degree distribution \( p_k \). The total number of stubs in our network is given by \( \sum_i k_i = 2E \), where
$k_i$ is the degree of node $i$ and $E$ is the number of edges in the network (equation 1.7). There are $2E - 1$ stubs excluding the stub starting at our initially infected node. Of these $2E - 1$ stubs, $k$ of them are attached to any node of degree $k$. The probability that the edge starting at the initially infected node ends on a specific degree-$k$ node is $\frac{k}{2E-1} \approx \frac{k}{2E}$. As previously seen in equation (1.18), if there are $n$ nodes in the network, there are $np_k$ nodes with degree $k$. Therefore, the probability that our initially infected node is attached to any degree-$k$ nodes is

\[
\left( \frac{k}{2E} \right) (np_k) = k p_k \langle k \rangle ,
\]

(1.40)

where $\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{2E}{n}$. Using this we can say that the probability of infection of a susceptible neighbour of degree-$k$ is

\[
\phi \left( \frac{k}{2E} \right) (np_k) = \phi \frac{k p_k}{\langle k \rangle} .
\]

(1.41)

Therefore, the expected number of vulnerable children nodes is

\[
\xi = \sum_k \phi \frac{k p_k}{\langle k \rangle} (k - 1) .
\]

(1.42)

If $\xi > 1$, we are in the super-critical regime and the tree grows - the infection spreads leading to an epidemic. If $\xi < 1$, we are in the sub-critical regime where the tree shrinks in time - the disease dies out resulting in no epidemic. In order to calculate the epidemic threshold we need to look at the transition point $\xi = 1$, where we go from the sub-critical to super-critical regime. Setting $\xi = 1$ in equation (1.42), we get

\[
1 = \sum_k \frac{(k^2 - k)p_k}{\langle k \rangle} \phi_c .
\]

(1.43)

Rearranging, we find the equation for the position of the epidemic threshold in the network

\[
\phi_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} .
\]

(1.44)

Written in terms of the fundamental parameters $\beta$ and $\tau$, we get

\[
\beta \tau = -\ln(1 - \phi_c) = \ln \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k^2 \rangle - 2 \langle k \rangle} .
\]

(1.45)

If $\beta \tau$ exceeds this value there is a possibility of a epidemic, though not the certainty depending on the location of the initial infected node in the network. If $\beta \tau$ is less than this value then an epidemic is impossible regardless of where the initial carrier is situated.

The SIR model on a network is equivalent to a percolation process - with $\phi$ being the fraction of occupied edges [36]. These occupied edges represent those along which the
disease will be transmitted if it reaches either of the nodes at the end of the edge. The
set of nodes to which the disease will spread is the set reachable from the initial infected
node via paths that go along occupied edges.

Unlike the ‘fully mixed’ model, in the network model when $\phi$ is above the epidemic
threshold we are not guaranteed an epidemic. Percolation is a stochastic process. The
occupied edges are chosen randomly for each realization. This translates to random
paths along which the disease can spread. Different realisations under the same condi-
tions can therefore lead to different nodes being involved in the epidemic and different
outbreak sizes depending on which edges are occupied and where the initially infected
node is located in the network. Due to this randomness, we are restricted and are only
able to calculate the average or expected fraction of individuals affected by an outbreak.

The epidemic behaviour is controlled by the model parameters $\beta$ and $\tau$. The epidemic
threshold can be controlled by the infectiousness time $\tau$, which is a property of the
disease under study or the transmission rate $\beta$, which is a property of the disease and
the behaviour of members of the population. The position of the epidemic threshold
is also dependent on the structure of the network via the moments $\langle k \rangle$ and $\langle k^2 \rangle$ of the
degree distribution. This contrasts with the ‘fully mixed’ version which incorporates no
network effects.

1.3.3 The SIS Model

1.3.3.1 The ‘Fully Mixed’ SIS Model

In this section, we briefly discuss another possible extension of the simple SI model. In
this new model, we allow for reinfection. That is an individual node can be infected more
than once. We call this the SIS model. As in the SI model, there are just two states in the
model, susceptible and infected. In this model, once a node recovers from the disease
it immediately goes back into the susceptible compartment, thus allowing for reinfection.

The fractions of $s$ and $i$ are governed by the equations

$$\frac{ds}{dt} = \gamma i - \beta si,$$

$$\frac{di}{dt} = \beta si - \gamma i,$$

where $s + i = 1$ [1]. We can solve for $i(t)$ to get the fraction of infected individuals
as a function of time. The solution produces a curve similar to that of the simple SI
model when $\beta > \gamma$ [1]. The difference here is we never have that the whole population
is infected with the disease. We can find the steady state of $i$ by setting $\frac{di}{dt} = 0$. This
gives $i = \frac{(\beta - \gamma)}{\beta}$. This is called the endemic disease state.
The fraction of infected in the endemic state is zero when $\beta = \gamma$. Thus as in the SIR model, $\frac{\beta}{\gamma} = 1$ marks a transition point between a state in which the disease spreads and one in which it dies out.

1.3.3.2 The SIS Model on a Network

As we have seen previously with the SI model and the SIR model, incorporating the network structure completely dictates the spreading dynamics. The SIS model on a network is the same. There is a new stochastic element involved when we consider the SIS dynamics taking place on a network. The outcome of the model is not only determined by the SIS model and its parameters, but also the structure of the network. Even in the case when $\beta > \gamma$, the disease may not spread to the $\frac{\beta - \gamma}{\beta}$ nodes as it does in the ‘fully mixed’ model. The overall outcome of the epidemic depends on the location of the initially infected node in the network and the pattern of connections between it and the rest of the nodes in the network.

1.4 Implementing Disease Spreading Models

In the previous section, we looked at various stochastic epidemic models on networks. A commonly used approach for investigating these dynamics is stochastic simulation [37]. It is normally the first step taken by researchers when studying the process of epidemic spreading models on a network.

In this section of the chapter, we introduce two different approaches that can be used when stochastically simulating epidemic models on a network.
1.4.1 The Independent Cascade Model

One way to study these epidemic models is by using discrete-time approximations. In a discrete-time approach, time is treated as a discrete variable of length $\tau$ - time is discretized into time steps of length $\tau$. In this framework, events occur with certain probabilities. In the SIR model, susceptible nodes become infected through their infected neighbours with a probability $\phi = 1 - e^{-\beta \tau}$ per infected neighbour, where $\beta$ is the transmission rate (equation (1.39)).

The Independent Cascade Model (ICM) is a discrete-time simulation scheme used to stochastically simulate SIR dynamics. In this model, time advances in time steps of length one, $\tau = 1$. In a single time interval a susceptible node can become infected by its infected neighbours with probability $\phi$ per infected neighbour, and its infected neighbour which passed the disease simultaneously recovers. The other possibility is that a susceptible node does not get infected by any of its infected neighbours in the time interval $\tau = 1$.

![Figure 1.10: The discrete-time process on a network. At time $t$ an infected node (red) has a susceptible neighbour (pink). At the next time step $t+1$, the susceptible node either becomes infected with probability $\phi$ and the infected node recovers (green) or the system remains unchanged.](image)

An important point to note is that the ICM has a synchronous updating scheme. A susceptible node becomes infected and the node that infected it simultaneously recovers.
in the $\tau = 1$ time interval.

### 1.4.2 The Gillespie Algorithm

Another approach is to study these epidemic models using continuous-time processes. Consider the SIR dynamics taking place on a network. In the continuous-time regime, susceptible nodes can become infected through each of their infected neighbours at a rate $\beta$ per infected neighbour, while the infected nodes recover at a rate $\gamma$. In continuous-time dynamics nodes change state one at a time. They are updated asynchronously.

![Figure 1.11: The continuous-time process on a network. At time $t$ an infected node (red) has a susceptible neighbour (pink). After a time $\tau$, either a susceptible node becomes infected which happens at a rate $\beta$ or an infected node recovers (green) which happens at a rate $\gamma$.](image)

The Gillespie algorithm is a well known algorithm for simulating continuous-time processes where objects change state [37]. The probability distributions governing the length of time until the next state change and which node will change state is given in the following Lemmas [38]:

**Lemma 1.1** Let $\tau$ be the length of time that the network remains in its current state before changing to a new state. Then $\tau$ is an exponentially distributed random variable and the parameter of the distribution is the total rate of change in the system.

**Lemma 1.2** The probability that the next node in the network to change state will be node $i$ is proportional to node $i$’s rate.
The algorithm finds the rates of all possible events and computes the total rate of change occurring. Using this total rate of change, the waiting time until the next event is chosen. Separately, an event is chosen with probability proportional to that events rate. The algorithm jumps to this time, updates states and corresponding rates and then repeats. Below is the pseudo code for the Gillespie algorithm.

**Algorithm 1** The Gillespie Algorithm.

**Input:** Network G, per-edge transmission rate \( \beta \), recovery rate \( \gamma \), Index of initially infected nodes, maximum time.

**Output:** List of times, S, I and R giving the number of nodes in each state at each time.

```
function Gillespie(G, \( \tau \), \( \gamma \), initial-infected, \( t_{max} \))
    times, S, I, R ← [0],[|G|-len(initial-infected)],[len(initial-infected)], [0]
    infected-nodes ← initial-infected
    at-risk-nodes ← uninfected nodes with infected neighbours
    for each node \( i \) in at-risk-nodes do
        infection-rate[\( i \)] = \( \tau \times \) number of infected neighbours
        total-infection-rate ← \( \sum_{i \in \text{at-risk-nodes}} \) infection-rate[\( i \)],
        total-recovery-rate ← \( \gamma \times \) len(infected-nodes)
        total-rate= total-infection-rate + total-recovery-rate
        time ← exponential-distribution(total-rate)
    while time < \( t_{max} \) and total-rate > 0 do
        r=uniform-random(0,total-rate)
        if r < total-recovery-rate then
            \( i \) = random.choice(infected-nodes)
            remove \( i \) from infected-nodes
            reduce infection-rate[\( j \)] for \( i \)'s susceptible neighbours \( j \)
        else
            choose \( i \) from at-risk-nodes with probability \( \frac{\text{infection-rate}[\( i \)]}{\text{total-infection-rate}} \)
            remove \( i \) from at-risk-nodes
            add \( i \) to infected-nodes
            for susceptible neighbours \( j \) of \( i \) do
                if \( j \) not in at-risk-nodes then
                    add \( j \) to at-risk-nodes
                update infection-rate[\( j \)]
            update times, S, I and R
            update total-recovery-rate, total-infection-rate, total-rate
        time ← time + exponential-distribution(total-rate)
    return times, S, I, R
```

### 1.4.3 Discrete-Time Versus Continuous-Time

In the previous sections, we considered discrete-time and continuous-time processes for simulating the SIR model on a network. Discrete-time approaches correspond to their continuous-time counterpart as \( \tau \to 0 \). However, both approaches can give completely different results when \( \tau \) is finite.
As previously discussed the ICM is a discrete-time process. The transmission probability per discrete-time step is $\phi$.

![The Independent Cascade Model](image)

Figure 1.12: The ICM process on a network used for simulating discrete-time dynamics.

Alternatively, the Gillespie algorithm is a continuous-time process. In this process, the transmission rate is $\beta$.

![The Gillespie Algorithm](image)

Figure 1.13: The Gillespie algorithm taking place on a network. This algorithm is used for simulating continuous-time dynamics.

Considering an infected node connected by an edge to a susceptible node in the SIR model. The probability of infection in the time interval $(t, t + \tau)$ is given by $1 - e^{\beta \tau}$. 
This is the transmission probability $\phi$ derived in Section 1.3.2.2. If $\beta \tau \ll 1$, then we can expand equation (1.39) and write

$$\phi = 1 - e^{\beta \tau} \simeq 1 - (1 - \beta \tau + \frac{\beta^2 \tau^2}{2} + \cdots) = \beta \tau + \frac{\beta^2 \tau^2}{2} + \cdots \simeq \beta \tau.$$ 

We can thus say that $\phi \simeq \beta \tau$ when $\beta \tau \ll 1$. We can compare the discrete-time transmission probability $\phi$ with the continuous-time transmission rate $\beta$ when the discrete-time step $\tau$ or the transmission rate $\beta$ are small. Thus if the discrete-time step or the transmission rate are small, the discrete and continuous-time processes can be compared.

Indeed, this connection between the transmission probability $\phi$ and the transmission rate and recovery time ($\beta$ and $\tau$) is useful. Such a relation would be helpful when trying to compare two models where the inputs given are the transmission probability in one (ICM) and the transmission rate and recovery rate (Gillespie Algorithm) in the other. As we can see in the Gillespie Algorithm pseudo code (Algorithm 1), the transmission rate and recovery rate are inputs into the algorithm and so having a connection between these two parameters and the transmission probability which is the input for the ICM could be useful. The transmission probability can be written in terms of the probability distributions of the transmission rate and the recovery time [39]

$$\phi = \langle \phi_{ij} \rangle = 1 - \int_0^\infty P(\beta)P(\tau)e^{-\beta \tau}d\beta d\tau. \quad (1.48)$$

Writing the probability distribution for $\tau$ using equation (1.37) (derived in Section 1.3.2.1)

$$P(\tau) = \gamma e^{-\gamma \tau}, \quad (1.49)$$

and since we are taking the transmission rate to be constant, we have that the probability distribution for $\beta$ to be given by

$$P(\beta_{ij}) = \delta(\beta_{ij} - \beta) \quad (1.50)$$

Then,
\[ \phi = 1 - \int_0^\infty \gamma e^{-\gamma \tau} e^{-\beta \tau} d\tau, \]
\[ = 1 - \int_0^\infty \gamma e^{-\tau(\gamma + \beta)} d\tau, \]
\[ = 1 - \frac{\gamma}{\gamma + \beta}. \]

Therefore
\[ \phi = \frac{\beta}{\gamma + \beta}. \] (1.51)

Later in this thesis, we will show results from simulations using the ICM and the Gillespie algorithm. We use the transmission probability \( \phi \) as the input for the ICM and the corresponding \( \beta \) and \( \gamma \) satisfying equation (1.51) as inputs for the Gillespie algorithm. This is a way to be able to compare and contrast dynamics occurring in discrete and continuous-time.

However, it is important to note that the dynamics cannot be directly compared. As we have previously discussed in this section, the ICM is a synchronous updating process and the Gillespie algorithm is an asynchronous updating process. Both models are fundamentally different in their dynamics. Nevertheless, they both simulate the SIR model and running them both with comparable parameters can support arguments if they both agree on a specific outcomes.
Chapter 2

Centrality

2.1 What is Centrality?

In principle, if we know the complete structure of a network, given the complete network data, then we know everything there is to know about that network. However, this network data is not easy to comprehend solely by inspection. For this reason, we develop mathematical measures. These measures are used to quantitatively capture features of the networks structure.

‘Who are the most influential nodes in a specific network?’ - this is one of the most popular questions in Network Science. Regardless of the specific area of research, we generally want to know which node is the cause of the biggest spread within the network, be it disease, rumours, information etc. We can take advantage of knowing which node is the most efficient spreader, for example, by targeting it to optimize the propagation of information or to immunize/isolate it in order reduce the extent of an epidemic.

Centrality measures have the purpose of quantitatively gauging the importance of individual nodes. A large portion of research on networks has been devoted to this concept of centrality. There are of course many possible definitions of importance, and therefore many different centrality measures for networks.

In this chapter, we briefly introduce the main centrality measures currently available. We highlight the advantages and disadvantages of the various measures. We then, define in detail a recently developed centrality measure called the Non-Backtracking (NB) centrality. We discuss some of the recent literature that investigates the NB matrix on which the NB centrality is based and discuss some of the possible advantages of using this matrix over the adjacency matrix of a network. We finish this chapter with a general literature review on the work to date surrounding the identification on influential nodes that we believe to be significant.

2.2 Centrality Measures

2.2.1 Degree Centrality

The most basic centrality measure is degree centrality [1]. This is just the degree of the individual node, the number of edges connected to it. In directed networks, a node will have an associated in-degree and out-degree. These can be useful when calculating the influence of individual nodes in directed networks.

Even though degree centrality is a very simple measure, it can be very useful and give some important insights into the specific network and nodes being investigated. For example, in social networks like Twitter, it would be reasonable to assume that a person
with many followers would have more influence than a person with fewer followers. Considering directed links from followees to followers, the followers are simply corresponding to the out-degree of the followee and hence give the out-degree centrality of that node. On the other hand, it’s plausible to suggest that a person with a high in-degree has a better spreading power than people with lower in-degree. For example, people with higher in-degree are exposed to a wider spectrum of information and may be more likely to receive and spread a piece of information. So a possible measure used to identify the ‘super-spreaders’ in the network could be the in-degree centrality.

2.2.2 Eigenvector Centrality

Degree centrality is a very basic centrality measure, however its computational simplicity comes at a cost. As already discussed, the degree centrality of a specific node is just that node’s degree. A node’s influence is simply just a measure of how many neighbours that node has. It can immediately become obvious that this might not necessarily be the case. If a node has important neighbours then that node will in essence be more important than a node with the same number of unimportant neighbours. A node’s importance is increased if it has important neighbours. In a sentence - It’s not just about how many people you know but also who you know. Degree centrality does not account for the importance of neighbours of a node in a network.

Eigenvector centrality [1] takes this issue into account. Eigenvector centrality gives each node a score proportional to the sum of the scores of its neighbours. The eigenvector centrality \( x_i \) of a node \( i \) is given by

\[
x_i = \lambda_1^{-1} \sum_j A_{ij} x_j.
\] (2.1)

Here the centrality of node \( i \) is the \( i \)th element \( x_i \), of the vector \( x \). The elements of the adjacency matrix \( A_{ij} \), makes sure the sum is only over all neighbours \( j \) of node \( i \). Here \( \lambda_1 \) is an eigenvalue of the adjacency matrix \( A \). From this equation it is clear that the eigenvector centrality of a node will be large either because a node has many neighbours or because it has important neighbours. This makes sense. You can be influential if you know a lot of people or know important people.

Eigenvector centrality can also be applied to directed networks. In these cases, the adjacency matrix is asymmetric and therefore has two sets of eigenvectors - the left and right eigenvectors. We generally use the right eigenvector as the measure of centrality. It makes sense that the importance of a node is due to the number of other nodes pointing to it. For example, a person’s influence on Instagram is dictated by how many people are following them. It is redundant how many people they follow back. For that reason we look at the right leading eigenvector as the centrality measure in directed networks.

However, there are still problems. A node in a network may be pointed to by many other nodes that themselves are pointed to by many more and so on. The issue arises if along this path of generations, we end on a node with zero in-degree [1]. This whole trail was for nothing and following equation (2.1) above, the in-degree centrality of the original
node is zero. We may have a very important follower, but if down the line of followers there is a node with no followers, then our influence is unaffected by having this original influential follower. In certain situations, this definitely would not make intuitive sense.

### 2.2.3 Katz Centrality

Katz centrality addresses the issue discussed above with eigenvector centrality [40]. Katz centrality is given by

\[
x_i = \alpha \sum_j A_{ij} x_j + \beta,
\]

where \( \alpha \) and \( \beta \) are constants. As a solution to the problem with eigenvector centrality, we give each node a small amount of centrality ‘for free’ regardless of its position in the network or the centrality of its neighbours.

Katz centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbours and also all other nodes in the network that connect to the node under consideration through these immediate neighbours. Connections made with distant neighbours are, however, penalized by an attenuation factor \( \alpha \). The first term in equation (2.2) is basically the eigenvector centrality and the second term is the constant extra amount given to all nodes.

By adding this constant term \( \beta \), we ensure that even nodes with zero in-degree get a constant amount of centrality regardless. This allows the centrality from important nodes to still be passed through the network. For this reason, Katz centrality supplies an answer to the issues that arose with eigenvector centrality in directed networks. It allows a node to have high centrality even if its neighbours don’t have high centrality themselves.

### 2.2.4 PageRank Centrality

Katz centrality [1] has one feature that can be undesirable. If a node with high Katz centrality has edges pointing to many others, then they also get high centrality. Sometimes this may be inappropriate. In many cases, such a centrality measure is less informative if a node is only among many that are pointed to; the centrality gained by virtue of receiving an edge from a prestigious node is diluted by being shared with so many others.

We can allow for this by defining a variation of Katz centrality, in which the centrality of a node derived from its network neighbours is proportional to their centrality divided by their out-degree. This is called Page-Rank centrality and is given by [41]

\[
x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_{out}^j} + \beta,
\]

Here \( \alpha \) and \( \beta \) are the same as in Katz centrality and \( k_{out}^j \) is the out-degree of node
PageRank centrality is the centrality measure used by the Google web search corporation. Google uses it to rank the importance of webpages for their web searches. PageRank centrality allows the Google search engine to list the most important websites first.

### 2.2.5 Closeness Centrality

Up to this point, we have looked into centrality measures based on the eigenvector of the adjacency matrix of the network. We will now examine two centrality measures that are based on the distance between nodes in a network.

Closeness centrality \cite{1} measures the mean distance from a node to other nodes. Closeness centrality is given by

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}},$$

(2.4)

where $d_{ij}$ is the length of the geodesic distance path from $i$ to $j$, meaning the number of edges along the shortest path from node $i$ to node $j$. Here $l_i$ is the mean shortest path between node $i$ and all other nodes. When $l_i$ takes a low value, we can consider the node in question to have a higher influence than if it had a higher value. For example, in a social network, it’s plausible to think that a node with a low mean shortest distance might find that their opinions spread throughout a community quicker than other nodes with a higher value.

This measure is different from others in the sense that more influential nodes have a lower score. For this reason, we take the inverse of it and equation (2.4) is defined to be the measure for closeness centrality.

A possible issue that arises with closeness centrality is when there are more than one component in the network being studied. In this case \cite{1}, if two nodes are in different components the shortest distance between them is infinite. The $l_i$ is infinite for all nodes and $C_i$ is zero for all nodes. A possible way to get around this problem is to redefine closeness centrality as

$$C_i = \frac{1}{n - 1} \sum_{j \neq i} \frac{1}{d_{ij}}.$$

(2.5)

Here, we average over the inverse distances instead. Here when any $d_{ij} = \infty$ (nodes in different components), the term just simply drops out. However, this definition of closeness centrality is rarely used by researchers in reality.

### 2.2.6 Betweenness Centrality

Betweenness centrality \cite{1} \cite{42} \cite{43} measures the extent to which a node lies on paths between other nodes. Let’s consider, a network with messages flowing through it, and
we make the assumption that the messages take the shortest path through the network and that all nodes exchange messages at the same average rate. A useful thing to know might be how many messages passed through each node en route to its destination. Since the messages take the shortest path, the number passing through each node will be proportional to the number of shortest paths the node lies on. Betweenness centrality is the number of these shortest paths.

Nodes with high betweenness centrality may have considerable influence within a network by virtue of their control over information passing between others. For example, if we target a specific node with high betweenness centrality and remove it from the network, we could then possibly reduce the spread of false rumours or news.

Betweenness centrality is given mathematically by

\[ x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}, \]  

(2.6)

where \( n_{st}^i \) is the number of geodesic paths from \( s \) to \( t \) that pass through \( i \), and \( g_{st} \) is the total number of geodesic paths from \( s \) to \( t \). We adopt the convention that \( \frac{n_{st}^i}{g_{st}} = 0 \) if \( n_{st}^i = 0 \) or \( g_{st} = 0 \). Unlike the other centrality measures we have looked at, betweenness centrality does not measure how well connected a node is in a network. Instead it is a measure on how much a node lies between other nodes.

### 2.2.7 K-Shell Decomposition

Another centrality measure to consider is \( k \)-shell decomposition [44]. Nodes are assigned to \( k \)-shells depending on their remaining degree. This is obtained by successive pruning of nodes with degree smaller than the \( k \) value of the current layer.

For example, we start by removing all nodes with degree \( k = 1 \). After removing all nodes with \( k = 1 \), some nodes may be left with one link, so we continue pruning the system until there is no node left with \( k = 1 \) in the network. The removed nodes, along with the corresponding links form a \( k \)-shell with index \( k_s = 1 \).

Each node is associated with one \( k_s \) index. The network can then be viewed as the union of all \( k \)-shells.

### 2.3 The Non-Backtracking Centrality

The final existing centrality measure that can be found in the literature that we introduce is the Non-Backtracking (NB) centrality [45]. NB centrality is an eigenvector based centrality. However, it does not use the leading eigenvector associated with the adjacency matrix. Instead it uses the leading eigenvector of another matrix called the Hashimoto matrix or just the NB matrix.
In the next section, we will describe in detail the structure of this matrix and understand its possible advantages over using the adjacency matrix. In the subsequent section, we introduce the definition of the NB centrality. We discuss some of the literature surrounding the topic in which the authors highlight its advantages.

2.3.1 The Non-Backtracking Matrix

The Hashimoto NB matrix is a representation of the link structure of a network that is an alternative to the usual adjacency matrix. The matrix is used to identify NB walks on a network. This means that the walks do not proceed from node \( i \) to node \( j \) only to immediately return to node \( i \). This property of the NB matrix makes it useful for defining centrality measures that correct for the potentially exaggerated importance given to high-degree nodes in common metrics like, for example, in eigenvector centrality.

Let \( G \) be an arbitrary and unweighted network with \( n \) nodes and \( E \) undirected edges. The NB matrix is a \( 2E \times 2E \) array, with one row and one column for each link in the network. The matrix is defined for directed links, but can easily be applied to undirected networks. In this case, each undirected link between \( i \) and \( j \) is replaced with two directed links \( i \rightarrow j \) and \( j \rightarrow i \).

The NB matrix encodes information about the sequences of links that we can follow in a walk through your network. More specifically, if we just traversed the link \( i \rightarrow j \), the NB matrix reveals what links \( k \rightarrow l \) are allowed in the next steps of the walk. It explicitly excludes the possibility of immediately backtracking from \( j \rightarrow i \).

The entries of the NB matrix is given by

\[
M_{i \rightarrow j, k \rightarrow l} = \delta_{j,k}(1 - \delta_{i,m})
\]

where \( \delta_{i,j} \) is the Kronecker symbol. \( M_{i \rightarrow j, l \rightarrow m} \) is different from zero and equal to one, only if the edges \( i \rightarrow j \) and \( l \rightarrow m \) define a NB path of length two. The NB matrix can therefore be expressed as

\[
M_{i \rightarrow j, k \rightarrow l} = \begin{cases} 
1 & \text{if } j = k \text{ and } i \neq l, \\
0 & \text{otherwise}.
\end{cases} 
\]

Visually the non-zero entries of \( M \) correspond to link pairs like links 1 and 3 in Figure 2.1 below. Meanwhile the NB condition sets entries to zero when links are oriented like 1 and 2.

The adjacency matrix for the network in Figure 2.1 below is given by

\[
A = \begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 0
\end{pmatrix},
\]

and the NB matrix is given by

\[
M
\]
We can now clearly see the difference between the adjacency matrix and the NB matrix for this directed network.

An interesting point to note is that powers of the NB matrix create walks through the network. Diagonal elements of the NB matrix correspond to closed walks that return to their starting point while fulfilling their NB condition. For example, if we multiply $M$ by itself three times we get

$$M^3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$ 

The matrix $M^3$ indicates that there are three closed NB walks of length three in the network. As can be seen in Figure 2.1 visually, there are three walks that begin at links 1, 3 and 4, and circle the triangle.

The NB matrix has been used in the area of community detection in networks [46]. It has been shown that the eigenvalues of the adjacency matrix that is correlated with community detection can fail to infer communities when homophily (birds of a feather flock together), in the network is weak. The reason is that the edges of the bulk of the spectrum are smeared by the influence of hub nodes, making it easier for the special, community correlated eigenvalue to get lost amidst the hub-dominated eigenvalues.
In [46], Krzala et al. show that a spectral algorithm based on the NB matrix can rectify this problem. There is an approximate eigenvector of the NB matrix that is highly correlated with the community-correlated eigenvector of the adjacency matrix. Using the NB matrix redeems spectral clustering because the NB property reduces the influence of hubs (nodes with high degree) [47] [48], and thereby reduces the smearing of the edges of the bulk spectrum.

Another benefit of using the NB matrix that has been discovered is the identification of the percolation threshold in tree-like networks. In [49], Karrer et al. show that we can get a good estimate of the percolation threshold \( \phi_c \), by taking the inverse of the leading eigenvalue of the NB matrix of the network. As we discuss in more detail in Chapter 3, they calculate the probability \( \pi_i(s) \), that a node \( i \) belongs to a finite cluster of size \( s \). This can be expressed in terms of the probabilities \( \pi_i \rightarrow j(s_j) \) that \( s_j \) nodes are reachable from \( i \) via the links \( i \rightarrow j \), for each of the potential neighbours \( j \) of \( i \).

In computing the number of nodes accessible from \( i \) in this way, the main approximation we need to make is that the structure of the network is tree-like, so that the same nodes are not reachable from two different neighbours \( j_1 \) and \( j_2 \). Continuing with this tree-approximation, \( \pi_i \rightarrow j(s_j) \) can itself be expressed in terms of \( \pi_j \rightarrow k(s_k) \), for each of the neighbours \( k \) of \( j \), excluding \( i \). The NB matrix, whose definition also considers links leading from a node to each of its neighbours but one, emerges naturally from this reasoning. Exploiting the stability of the fixed point of their eigenvalue equation, they are able to deduce that the percolation threshold is simply the inverse of the leading eigenvalue of the NB matrix.

Although the approximations used by Karrer et al. are strictly valid for only tree-like networks, the authors show numerically that, in practice the inverse of the leading eigenvalue of the NB matrix can be used to identify the percolation threshold in real-world networks.

In this section, we have introduced the NB matrix and given examples where it can unfold properties of networks, where the adjacency matrix fails to do so.

### 2.3.2 The Non-Backtracking Centrality

Before giving the definition of the NB centrality, we briefly revise the mathematics of eigenvector centrality. As previously discussed, this measure is based on the intuition that a node’s importance in a network is determined by how many neighbours it has and how important its neighbours are in the network. We can initialize every node’s centrality to one, and then iteratively update the centrality of each node to be the sum of their neighbours centralities.

For an undirected network, these iterations are given by

\[
x_{m+1} = Ax_m.
\]

where \( A \) is the adjacency matrix, which encodes the neighbours of each node \( i \). Here \( x_m \) is the \( m \)-dimensional vector with the centrality estimates for each node after the \( m^{th} \)}
iteration for a network of size \( n \).

Repeating this iterative process, eventually the relative weights of the centrality measures in \( x_m \) will converge to the leading eigenvector of the adjacency matrix \( A \).

Looking at eigenvector centrality from first principles, like we have above, helps display the issues that arise with it. A central node in a network will pass its centrality along to its neighbours, and then, in the next step, that centrality gets reflected back to that central node. This repeated reflection can lead in unjustifiably large accumulation of eigenvector centrality near high degree nodes in a network. We call this phenomenon localization.

A possible solution to this problem [45] is to find the leading eigenvector \( \epsilon_{i \rightarrow j} \) of the NB matrix instead. The element of this eigenvector corresponding to the link \( i \rightarrow j \), can be interpreted as an eigenvector centrality of node \( j \) that neglects any contribution from node \( i \). With this eigenvector, we can compute the full centrality of node \( j \), by summing over all neighbours \( k \) of \( j \). Therefore, the NB centrality of a node \( i \) is given by

\[
x_i = \sum_j A_{ij} \epsilon_{i \rightarrow j},
\]

where \( \epsilon_{i \rightarrow j} \) is the leading eigenvector of \( M_{i \rightarrow j, l \rightarrow m} \), and \( A_{ij} \) are the elements of the adjacency matrix of the network. This appears to actually capture the intuition we hope to capture in eigenvector centrality. When we ask what importance our neighbours pass onto us, we really want to consider how important our neighbours would be in our absence.

### 2.4 A Review of the Literature: Identification of Influential Nodes

We dedicate this section to a discussion of some of the research and work that we believe is significant to the area of influential node identification on a network. We look at the literature that is available in which different centrality measures are considered for the identification of ‘super-spreaders’. Throughout this literature review, we compare and contrast the different methods and views that have been put forward to date when trying to single out or rank the most efficient spreaders in a network.

In [44] the authors claim that there are circumstances where the best spreaders in a network do not necessarily correspond to the most highly connected or most central people. Instead, it is found that the most efficient spreaders are those located within the core of the network, as identified by the \( k \)-shell decomposition analysis. The authors of [44] suggest that when multiple spreaders are considered simultaneously, the distance between them becomes a crucial parameter that determines the extent of the spreading. It is also shown that infections persist in the high \( k \)-shells of a network, in the case where recovered individuals do not develop immunity (The SIS model).
In the case of networks with a broad degree distribution, it is well-known that the nodes with largest degree are the key players, responsible for the greatest extent of spreading. In social networks, the importance of a node for spreading is often associated with betweenness centrality. However, in [44] it is argued that the topology of the network plays an important role such that there are circumstances under which nodes with largest degree or the highest betweenness nodes have little effect on the range of a given spreading process.

In [44] a $k$-shell decomposition of the network is used to identify the core and periphery of the network and sort all of the nodes into various $k$-shells. The authors make use of real-world complex networks that represent archetypical examples of social structures. The SIR and SIS models are applied on the above networks to study the spreading process. Specifically, a small value for the infectious rate ($\beta$) is used. This is to ensure the infected percentage of the population remains small and the role of the individual nodes remains important.

Using the contact network of inpatients collected from hospitals in Sweden, the authors of [44] show that the size of the population infected is not necessarily related to the degree of the node $k$, where the spreading started. Instead, it is shown that the spreading origin given by its $k_s$ index predicts more accurately the size of the infected population. Nodes in the same $k_s$ layer produce similar spreading areas even if they have different degree. To quantify the influence of a given node $i$ in an SIR spreading process, the average size of the population $M_i$ infected in an epidemic originating at node $i$, with given ($k_s, k$) is studied. The infected population is averaged over all the origins (seed nodes) with the same ($k_s, k$) value ($M(k_s, k)$). It is found that for fixed degree, there is a wide spread of $M(k_s, k)$ values. For fixed $k_s$, $M(k_s, k)$ is found to be approximately independent of the degree of the nodes and that the efficient spreaders are located in the inner-core of the network, fairly independent of their degree. Similar results are obtained from the analysis of $M(k_s, C_B)$, where $C_B$ is the betweenness centrality of a node.

The imprecision functions, $\epsilon_{k_s}, \epsilon_k$ and $\epsilon_{C_B}$ (mathematical definition given in Chapter 3) is used to quantify the importance of $k_s$ in spreading. These functions estimate for each of the three indicators $k_s$, $k$ and $C_B$, how close to the optimal spreading is the average spreading of $\rho n$ ($0 < \rho < 1$) chosen origins (seed nodes) in each case. It is found that, using $k_s$ to predict the most efficient spreading nodes is consistently more accurate in the studied $\rho$ range. In the case of collective spreading (epidemic that starts at multiple origins simultaneously), the nodes with highest degree are more efficient than those with highest $k_s$. However, if the nodes aren’t directly linked, then both highest $k$ or $k_s$ nodes yield a similar result.

In the SIS model, it is found that $p_i(t \to \infty)$ (the probability that a node $i$ is infected at time $t \to \infty$), is largest in high $k_s$ layers in real-world networks. It is found that viruses persist mainly in high $k_s$ layers even when the infection rate is below the threshold ($\beta < \beta_c$). In both SIR and SIS models, the persistence $p_i$ or the average infected fraction $M_i$ is larger for nodes in inner $k$-shells compared with nodes in outer $k$-shells, over the entire $\beta$ range studied.

In [50] it is also acknowledged that degree centrality is of little relevance in many sit-
uations on real-world complex networks. The authors of [55] agree that betweenness centrality can be a good measure for the identification of influential nodes. It is also mentioned that closeness centrality is a very accurate centrality measure in certain situations. However, the authors acknowledge that for large scale networks this can be compromised due to computational complexity. The authors propose a ‘semi-local’ centrality measure for undirected networks, as a trade-off between the low-relevant degree centrality and other time consuming measures. The local centrality $C_L(v)$ of a node $v$ is defined as,

$$C_L(v) = \sum_{u \in \Gamma(v)} Q(u),$$

(2.13)

where $Q(u)$ is given by,

$$Q(u) = \sum_{w \in \Gamma(u)} N(w),$$

(2.14)

and $\Gamma(u)$ is the set of the nearest neighbours of node $u$ and $N(w)$ is the number of nearest and next neighbours of node $w$. As in [49] the SIR model is used to examine the spreading influence on nodes ranked by different centrality measures. Simulations on four real-world networks (Blogs, Routers, Nnetscience and Email) are run, to show that this ‘semi-local’ centrality can well identify influential nodes.

The total number of infected nodes and recovered nodes at time $t$ (in this paper, they let $t=10$), denoted by $F(t)$ is considered here as the indicator to evaluate the influence of the initially infected node. By testing the correlation between spreading influence $F(t)$ and the centralities, it is shown that local and closeness centrality measures perform much better than degree and betweenness centrality. It is noted that the performance of local centrality depends on the network structure. According to the authors, the local centrality is more suitable to be applied to heterogeneous networks. However, even in a tree-structure network, the local method can successfully show which nodes are more influential. Generally (over the four different networks), it is shown that local centrality has the strongest correlation with closeness centrality and weakest with betweenness centrality.

In [51] a physically grounded method is presented which solves the problem of identifying the most influential nodes exactly and performs very well in a very broad spectrum of situations. The work in this paper is based on the connection between bond percolation and the properties of the SIR model for epidemics. The authors of [51] highlight the role played by the spectral properties of the NB matrix in determining the properties of the bond percolation process in complex networks.

To model the spreading dynamics, the SIR model is used again. The total number of nodes whose final state is $R$ is denoted by $Q_t$, it represents the extent of the spreading event originated by the single node $i$. Similarly to [44] and [50] the SIR dynamical process is simulated with a single initial node $i$ in state $I$ and all other nodes in state $S$. Once the dynamics end, the number of $Q_t$ nodes in state $R$ is recorded, and the procedure is repeated $10^4$ times. The spreading power of node $i$ is quantified as $\langle Q_t \rangle$. As in [44] and [50] the measure $\langle Q_t \rangle$ and its associated ranking is used as the benchmark to which
they compare the centralities proposed to identify the influential nodes. Here, degree, $k$-core index, eigenvector and generalised random walk accessibility (RWA) centrality measures are considered.

The imprecision function $\epsilon$ (spreading power) and the Jaccard distance $d_J$ (to the ground truth ranking of the nodes) (mathematical definition given in Chapter 3) are used to compare the performance among the centrality metrics. Both measures take as input two sets of nodes. The first is the list of $\rho n$ ($0 < \rho \leq 1$) actual top-spreaders from the numerical simulations of the SIR model. The second is the list of $\rho n$ top nodes when ranked according to the centrality score $x_i$. Both the imprecision function and the Jaccard distance, return a value ranging between zero (for perfect matching) and one (for completely failed prediction).

In [51] it is noted that the percolation threshold in locally tree-like networks is exactly given by the inverse of the largest eigenvalue of the NB matrix. The work of [49] is cited and the authors of [51] explain that the probability of a node $i$ being part of the percolating cluster immediately above the threshold is given by,

$$x_i = \sum_j A_{ij} \epsilon_{i\to j},$$

where $\epsilon_{i\to j}$ is the leading eigenvector of the NB matrix and $A_{ij}$ are the elements of the adjacency matrix (we prove and explain this later in Chapter 3). The authors point out that the relative size of an epidemic outbreak originating from a node $i$ is proportional to the probability that $i$ belongs to the percolating cluster. At the critical point, this probability coincides with the NB centrality, thus

$$\langle Q_i \rangle = x_i.$$  (2.16)

This is an exact result, provided that the network structure is tree-like. It is noted that, however this is only relevant at criticality. The authors do however, argue that it is only around criticality that the choice of initiator will have a substantial impact on the spreading events.

In [51] the various centralities are compared on tree-like synthetic networks. They consider a distribution $P(k) \sim k^{-\gamma}$, with $\gamma = 3.5$. It is concluded that the NB centrality is the optimal choice for the selection of the influential spreaders in locally tree-like networks at criticality. A very large collection of real-world topologies (not considered tree-like) is also considered. Here, it is shown that eigenvector and NB centralities were very effective in identifying influential spreaders. From the analysis, it is found that $k$-shell centrality generally provides unsatisfactory results, compared not only with NB centrality but also degree, eigenvector and RWA centrality. This is at odds with what was claimed in [44]. It is claimed in [51] that their message is conclusive. The authors of [51] suggest that the large degeneracy involved in the $k$-core index would explain its poor ability to identify top-spreaders.
The authors of [51] mention that another exciting line of research regards the identification of influential spreaders from empirical data on real-world spreading phenomena. In this respect, the problem is further complicated by the fact that the spreading dynamics at the microscopic level are not known \textit{a-priori} and may contain additional ingredients not included in the simple models usually considered.

In [52] the authors attempt to unravel the problem posed in [51]. It is acknowledged that a number of predictors have been suggested to detect the most influential spreaders of information in online social media across various domains such as Twitter and Facebook. However so far, validation of the proposed predictors has been done by simulating basic spreading dynamics rather than following real information in social networks. Therefore, only model dependent contradictory results have been achieved so far for the best predictor. This issue is addressed directly and the influential spreaders are detected by following the spreading dynamics in a wide range of networks. From [44] [50] and [51] we find that searching for individual ‘super-spreaders’ of information is commonly implemented by ranking the users in terms of topological measures. However, so far there has been no consensus on the best predictor of influence.

A number of different measures aimed at identifying influential spreaders were suggested over the years. In [52] the authors point out that unavailability of the full diffusion record in networks has prevented straightforward validation of the efficiency of such measures. The drawback of previous studies is that the validation of the proposed predictors has been done by modelling the spreading of the information in a given network, rather than using the real spreading dynamics. This has led to a number of papers with contradictory results on the best predictor of influence like we have seen in [44] and [51].

The models used in [44] [50] and [51] are typically based on very simplified assumptions of human behaviour, that may not be representative of the actual information spreading dynamics in a real setting. For this reason, they give rise to model-dependent predictors for the best spreaders. The models usually fail to account for key elements affecting information consumption like user activity, individual interests and the distribution of these properties in the network. These issues motivate the authors of [52], to empirically test the variety of suggested predictors of influence using real information diffusion dynamics to find practical and reliable topological identifiers of super-spreaders of information.

The full empirical investigation of super-spreaders of information is investigated by following the real diffusion dynamics in some important online social networks. Spreading in such systems can be thought of in terms of two layers; the underlying social network and the diffusion process embedded in the population. The full information dynamics and topological network structure of a large data set representing public blog posts published at LiveJournal.com (LJ) is collected. In LJ, each user maintains a friend list, which represents social ties to other LJ users. It is claimed that the network composed of these social links is a reliable representation of the actual social relations of the LJ users. The friends of all users and all the available blog posts from February 14\textsuperscript{th} 2011 to November 21\textsuperscript{st} 2011 were recorded. From this it was possible to track the information passed from one user to another. In total 598,833 posts that contain links to other posts published by LJ users were identified and a diffusion link from \( j \) to \( i \) was defined if \( i \) cites \( j \)’s blog at least once. From this, a directed unweighted diffusion graph representing
information spread in LJ during the observation period was obtained. It is suggested that this complete network structure enables them to test the various centrality measures accurately.

It is found that the diffusion graph is quite different from the underlying social network. The diffusion graph is relatively small compared to the size of the underlying social network. It is also found that there are situations where information spreads between two users even if they are not connected by a social link. In [52] the diffusion links starting from each node \( i \) in LJ are followed. The resulting set of nodes from the diffusion cascade, represent the region of influence for node \( i \). The impact of the node \( i \), to the information spreading process is quantified as the number of nodes \( M_i \) in the region of influence. It is found that the location of the origin given by its \( k_s \)-index predicts the influence more accurately than degree, in-degree or PageRank centrality. It is claimed that \( k_s \) can predict the average influence well since the influence for single nodes has large fluctuations and that \( k_s \) is the better measure for identifying super-spreaders.

The dissemination of scientific information in the publication of the APS Journals is also explored. The dataset includes information of authors and citations for all publications until 2004. The social network is formed by co-authorship. The diffusion of information is reflected by citations. Although the mechanism of scientific ideas is different from that of LJ, in both cases, \( k \)-core outperforms degree and PageRank.

The authors of [52] acknowledge that unlike LJ and APS, we usually are unable to obtain the data of the complete social network. For this reason, it would be desirable to identify spreaders for networks where we do not have the complete network structure. In order to check the performance of \( k \)-core in networks with partial links, the subnetworks sampled from Facebook and Twitter are analysed.

For the incomplete Facebook network, it is found that \( k \)-core outperforms in-degree and PageRank. This is consistent with LJ and APS. For Twitter, a partial mention network (@username) is used. It is found that \( k \)-core is the superior measure. The authors of [52] claim that \( k_s \) index is a reliable predictor of influential spreaders due to the fact that measurements were taken on diverse datasets. The authors explain that the evaluation of \( k_s \) in real-world scenarios is frequently infeasible, (it’s a global measure). They introduce a simple local measure given by,

\[
k_{\text{sum}}(i) = \sum_{j \in V(i)} k_j,
\]

where \( V(i) \) is the set of nearest neighbours of the node \( i \). It is found that \( k_{\text{sum}} \) works quite well and can be used to identify the best spreaders, even though it uses partial information. These conclusions are again at odds with the findings in [51], where it is suggested that \( k \)-shell centrality is an unsatisfactory measure. The degeneracy inherent in the \( k \)-shell centrality must be acknowledged when interpreting the above results.

So far, we have referenced literature in which a multitude of centrality measures are compared on various spreading models. We can see from these examples that global
centrality measures such as betweenness, closeness, Katz, k-shell decomposition, eigenvector, NB centrality etc. can be very successful in identifying structurally important nodes. However, in [53] the authors suggest that these measures fail to identify nodes that are key players in dynamical processes taking place on the network.

The authors of [53] present a new centrality measure based on a stochastic spreading process, the transmission centrality, that captures the importance of edges by estimating the average number of nodes to whom they transfer information during a global spreading process. It measures the average number of nodes who are reached by the spreading process through each edge during the stochastic process. This gives a direct measure of the centrality of each edge in the network. It is claimed that this measure can be used to identify the key nodes and edges responsible for controlling the spread of information on a network.

The centrality measure aims to measure for each edge in a network its influence in the dissemination of some globally spreading information. The SI model is used to simulate the spreading process. To start with, a random initial seed node is chosen. During the process, the infected nodes with susceptible neighbours are recorded and the branching tree rooted from that seed node is calculated. For each leaf edge in the branching tree, its counter is increased by 1. All leaves are removed and the counter of their ascendants is increased by the counter of the removed leaves. The transmission centrality of an edge can takes values from $0$ and $n - 1$, depending on the seed node and the structure of the branching tree determined by the stochastic spreading process. To eliminate this effect, the average centrality value for each edge is taken and computed over processes initiated from every single node in the network. This is an edge-based centrality measure but can easily be defined as a node centrality by counting for each node the number of descendant nodes in the branching tree.

In [53] the properties of the measure are demonstrated on three large scale networks. A dataset is collected from mobile call communication sequences of 1,926,787 individuals during a 4-week period. Moreover, a social network aggregated from a sequence of posts from 20,244 Facebook users from September 2004 to January 2009 and also a Twitter conversation network of 966,779 users constructed using tweets from October 2010 to November 2013 are considered.

Zhang et al. [53] note at the end of this paper that their aim is to contribute to the design of a method used to identify the edges that play an important role in dynamics of spreading processes taking place on social networks.

Before we conclude this section, we would like to refer to the work in [54]. This is a review paper in which the authors discuss the current work surrounding information cascades on complex networks, with an emphasis on the role of node centrality. The authors begin the review with an introduction to the area of centrality and information cascades on networks and reference some key work that has been undertaken in this area.

The authors of [54] define models for information spreading and cascades on networks, including the independent cascade model (used in the work of this thesis). They refer to the influence maximization problem that has been discussed in the literature. This is
the question of how to choose the initially infected nodes. This is an important issue in the studies related to information cascades. We may want to know what nodes we need to initially infect in order to maximize the overall epidemic outbreak. On the other hand, we may want to know which nodes to isolate in order to avoid and minimize an outbreak. Simulation results are presented on sample networks that reveal how relevant the centrality of the initiator node is on the subsequent development of the information cascade. It is shown from these results that there are centrality measures positively correlated with the spreading influence and that there are also centrality measures negatively correlated with the spreading influence.

In [54] Jalili et al. dedicate an extensive portion of this review to the applications of these results. They do this to emphasize the importance of information cascades in helping us to identify nodes with the highest spreading capabilities. Some application areas they include are the mitigation of contagious disease and preventing epidemics, viral marketing, opinion spreading and consensus, prevention of cascading failures in infrastructure networks and contagion in financial systems.

The objective of this section is to demonstrate, discuss and compare the work that has previously been undertaken in this area. We can now clearly recognise that there are several centrality measures in use that have been tested on various network structures with various spreading models.

We have discussed a range of centrality measures that currently appear in the literature. Something to note is that most of the centrality measures we have discussed to this point are structurally dependent - the measures depend on the structure of the network. The centrality of a node is controlled by its location in the network. In the following chapter, we propose a new centrality measure that takes the structure and the dynamical processes taking place on the network into account.
3.1 Classical Branching Processes

In this section, we review some of the standard tools and concepts used when dealing with branching processes [55] [56]. This is a broad area of mathematics, so we only consider the basics that are necessary for the complete understanding of work later in the thesis. We use this theory when we introduce a new centrality measure, which is based on a branching process approach.

3.1.1 Probability Generating Functions

A probability generating function of a random variable is a power series representation of the probability mass function of the random variable. Let \( p_k \) be the probability that a random variable takes the value \( k = 0, 1, \ldots \); then, the probability generating function (PGF) associated to this random variable is

\[
 f(x) = \sum_{k \geq 0} p_k x^k. \tag{3.1}
\]

Some trivial but important properties are:

1. \( f(1) = \sum p_k = 1 \).
2. \( f'(1) = \sum kp_k = \langle k \rangle = E[k] \).
3. \( f''(1) = \langle k^2 \rangle - \langle k \rangle \Rightarrow \text{Var}(k) = f''(1) + f'(1) - [f'(1)]^2 \).
4. Let there be two random variables with distributions \( p_k, q_k \) and associated PGFs \( f(x), g(x) \), then the probability that the sum of the variables takes the value \( k \) is

\[
 \sum_{j=0}^{k} p_j q_{k-j} = \sum_{j=0}^{k} p_{k-j} q_j \Rightarrow f(x)g(x) = \sum_{j \geq 0} p_j x^j \sum_{l \geq 0} q_l x^l.
\]

This property tells us that the PGF of the sum of variables is the product of the PGFs.
3.1.2 Galton-Watson Branching Processes

A branching process is a stochastic process. It is a mathematical model of population growth in which each individual in generation \( n \) produces a number of individuals in the next generation \( n + 1 \), according to some probability distribution.

Some of the earliest work surrounding branching processes was derived in the mid 19\(^{th}\) century by two men, Sir Francis Galton and Henry William Watson. They used branching processes to investigate the extinction of family names. A commonly used formulation for branching processes is the Galton-Watson Process [57] [58]. In a Galton-Watson process, there is one individual at generation \( n = 0 \). The life duration of a particle (individual) is one, it then dies and spawns a certain amount of children with some probability.

When we have a system that can be modelled by a branching process, we can answer some interesting questions about that system being studied. In our case, we are interested in knowing the number of particles alive at generation \( n \) spawned from an initial active parent. The reasons we are interested in this will become clear in the next section.

We define \( G(\tau, \Omega; x) \) as the PGF for the size of the tree. The tree size is the number of children, \( \tau \) is the seed time, \( \Omega \) is the observation time [59]. Using the Galton-Watson branching process framework, we start with one parent node as can be seen in Figure 3.1 below at the seed time \( \tau \). This parent node then spawns a certain amount of children with some probability following the offspring distribution at time \( \tau + 1 \), and simultaneously dies. This process continues with the new parent nodes spawning further children according to the specified offspring distribution until the observation time \( \Omega \). In the example shown in Figure 3.1, the parent node spawns two children at time \( \tau + 1 \) and then they go on to spawn further children dictated by a probability distribution which is represented by the blue boxes in the schematic. The tree size is the total number of nodes (dashed pink box) that have been alive by the time of observation \( \Omega \).

We can write \( G(\tau, \Omega; x) \), which is the PGF for the tree size created by a parent at time \( \tau \) in terms of \( G(\tau + 1, \Omega; x) \) which is the PGF for the tree size created by a parent at time \( \tau + 1 \). Knowing the offspring distribution \( f \) of the system and accounting for the initial parent node at time \( \tau \), we can write [59]

\[
G(\tau, \Omega; x) = xf(G(\tau + 1, \Omega; x)),
\]

\[
\implies G_n(\tau, \Omega; x) = xf(G_{n-1}(\tau + 1, \Omega; x)), \quad (3.2)
\]

where \( G_0(x) = x \) following the PGF definition in equation (3.1) and \( n = \Omega - \tau \) is the tree age.
A branching process can be defined by a quantity $\xi$ called the branching number of the process. This is the mean number of offspring per parent. Depending on its branching number, the process can be classified as follows:

- $\xi < 1$ sub-critical,
- $\xi = 1$ critical,
- or
- $\xi > 1$ super-critical.

In the sub-critical regime with $\xi < 1$, the expected number of children per parent is less than one and so in the limit of large time the size of the tree is finite. When $\xi = 1$, we call this a branching process in the critical regime. In the super-critical regime, $\xi > 1$ and so the expected number of children per parent is greater than one. In the limit of large time, the tree size tends to infinity.

In the next section, we introduce a new centrality measure based on a branching process approach. We want to be able to predict the top influential nodes in a network. We think of the infection of individuals in an epidemic spreading as a spawning procedure in a branching process and define the most efficient spreaders to be the seeds responsible for the biggest tree (outbreak) size in the steady state (Section 1.4). We wait for the branching process to finish, and calculate the extent of the outbreak, be that it takes over a large fraction of the network or it dies out completely. Our new centrality measure is based on a branching process model and so the above result is somewhat of interest. We will see from the theory that in the limit of large age the branching process approach only returns meaningful results up to the critical regime.

### 3.2 A Mathematical Model: Predicting the Most Influential Nodes in a Network

In this section, we propose a new centrality measure based on a discrete-time branching process model. Given a network $G$ with $n$ nodes and $E$ undirected edges, we want to be
able to rank the nodes in the network from most influential to least influential. From the given network, we create a directed network where each undirected edge in the original network becomes two directed edges.

In order to determine the influence of a node $i$, we calculate $m_{i \rightarrow j}(a)$, the expected size of a tree (infection) at age $a$, created by a transmission attempt from $i$ to $j$ ($i \rightarrow j$). Summing over all $m_{i \rightarrow j}(a)$, where $j$ denotes the nearest neighbours of node $i$, and including $i$ itself, we can calculate the tree size created from seed $i$ at age $a = \Omega - \tau$.

Using the discrete-time branching process approach, we can write an equation for $m_{i \rightarrow j}(a)$. As previously shown in equation (3.2), this approach allows us to write the expected size of an outbreak created by a transmission along the edge $i \rightarrow j$ in terms of the expected sizes of outbreaks created by later transmissions. We can therefore write $m_{i \rightarrow j}(a)$ in terms of the sum of $m_{j \rightarrow k}(a - 1)$, where $k$ denotes the neighbours of the $j$ node. Including the initial node $i$, and taking into account the transmission probabilities along the edges, we can write the equation for $m_{i \rightarrow j}(a)$ as

$$m_{i \rightarrow j}(a) = \phi_{i \rightarrow j} \left( 1 + \sum_{k \in N(j) \setminus \{i\}} m_{j \rightarrow k}(a - 1) \right).$$  \hspace{1cm} (3.3)

This resembles a branching process algorithm which we introduced earlier in Section 3.1.2. Here, we have $m_{i \rightarrow j}(0) = 1$ as the initial condition and $\phi_{i \rightarrow j}$ is the transmission probability for the edge $i \rightarrow j$, and $N(j)$ denotes the set of neighbours of node $j$. 

Figure 3.2: One directed edge becomes two directed edges.
By using a branching process approach to model the spread along the network, we are assuming that the children of each node are independent from each other. This can be seen in Figure 3.3 above. The neighbours of $j$ ($k$ nodes) are assumed to not be neighbours of each other. This is an approximation which is a consequence of our branching process approach. However, we propose that this approach will be very accurate on networks that are locally tree-like (Section 1.2.1), that is networks with no short loops such as triangles. Since there are $E$ edges in the undirected network, equation (3.3) is a system of $2E$ equations.

Figure 3.3: A schematic of the branching process approach used to calculate the tree size created by a transmission along the $i \rightarrow j$ edge. Here $0$ and $a$ represent the system times.

Figure 3.4: A schematic of the branching process approach used to calculate the influence of node $i$. Here $0$ and $a$ represent the system times.
In order to calculate the total influence of node \( i \), we need to sum the expected tree sizes resulting from all initial transmissions starting at node \( i \). Therefore, we can write the expected tree size initiated by node \( i \) at age \( a \) as

\[
S_i(a) = \sum_{j \in N(i)} m_{i \rightarrow j}(a) + 1. \tag{3.4}
\]

The tree size from node \( i \) is 1 (for node \( i \) itself) plus the sum over all tree sizes that were a result of a transmission from node \( i \) to all of its nearest neighbours (Figure 3.4). We repeat this process for every node in the network acting as the seed and rank the nodes according to their corresponding tree size from largest to smallest. We use this as a measure of importance of the nodes and hence propose this branching process model score as a centrality measure.

### 3.3 Implementation of the Branching Process Centrality Measure

In order to implement the branching process (BP) centrality, we use an iterative method. We start with an initial vector \( m_{\text{old}} \), where each element corresponds to an edge in the network. Every element in this vector is the size of the tree created by a transmission on the corresponding edge. We set the initial value of each element to one. We use the NB matrix to implement equation (3.3). It ensures that there are no backtracking steps during the spread of infection and so models the branching process algorithm. For the implementation, we use the matrix equation

\[
m_{\text{new}} = \phi(I + Mm_{\text{old}}). \tag{3.5}
\]

Here \( M \) is the NB matrix of the network, \( I \) is the identity matrix and \( \phi \) is the transmission probability (assuming it the same over all edges here). We replace \( m_{\text{old}} \) with \( m_{\text{new}} \) at the end of every iteration and repeat the process.

As previously mentioned, we rank the nodes based on the extent of the outbreak in the steady state. For this reason when implementing the centrality measure in equation (3.3), we consider the case where we take the limit \( a \rightarrow \infty \). As we saw in Section 3.1.2, in the large age limit, the expected size of a tree is finite in the sub-critical regime. This is when the expected number of children per parent is less than one. We propose the BP centrality to be suitable in this regime. Processes in the sub-critical regime can be estimated to follow tree-like patterns in their dynamics and so are not likely to violate the assumptions of the BP centrality measure. We consider a small transmission probability \( \phi \), that ensures the process is in the sub-critical regime with \( \xi < 1 \). Recalling equation (1.42) from Section 1.3.2.2 that

\[
\xi = \sum_k \phi^k \frac{k p}{k} (k - 1).
\]
Therefore, in order to be in the sub-critical regime, we need $\xi < 1$, and so

$$\phi < \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} = \phi_c.$$  \hspace{1cm} (3.6)

We call $\phi_c$ the critical transmission probability. We iterate equation (3.5) until the largest difference between the corresponding elements of $m_{\text{new}}$ and $m_{\text{old}}$ is less than a predefined tolerance and we reach a steady state, which we denote by $m_\infty$.

In the calculations performed for the work in this thesis, we set the tolerance of convergence to be $10^{-11}$. We test the outcomes on a few examples using a tolerance of convergence as small as $10^{-16}$. The difference in the resulting $m_\infty$ vectors is negligible and doesn’t alter the node rankings when comparing both tolerances of convergence on the same network. We therefore consider $10^{-11}$ to be a satisfactory tolerance and use this in all of our examples.

This iterative approach is not the only method we can use when calculating the BP centrality. In the steady state, we can write equation (3.5) as

$$m_\infty = \phi (I + M m_\infty).$$  \hspace{1cm} (3.7)

Rearranging this equation, we get the linear system

$$\left( I - \phi M \right) m_\infty = \phi I.$$  \hspace{1cm} (3.8)

In theory, this equation can be solved by finding the inverse of the matrix $I - \phi M$. However, when we deal with large networks the computational efficiency of these calculations are very poor and sometimes even just the storage of such large matrices causes issues. In Section 3.4.4, we show an example where we compare the computational efficiency of the iterative approach and the method of solving the linear system. Due to its computational superiority, the BP centrality is calculated using the iterative approach for all of the examples considered in this thesis.

Once we know the $m_\infty$ vector, there is then one more step in calculating the BP centrality. Each element in this steady state vector $m_\infty$ represents the sizes of trees created by a transmission along the corresponding edges. We then sum all contributions of edges starting with the same seed and add one to account for the seed itself. The result is an $n \times 1$ vector, where each element $i$ is the tree size created from the seed node $i$. We use this as a centrality metric and rank nodes according to their value in this vector.

Later in the thesis, we also implement the BP centrality in the critical and super-critical regime to test its performance. From the theory, we know that the BP centrality will be less accurate in these regimes due to the presence of short loops that may be in the network. In these cases, iteration of equation (3.5) doesn’t converge. In order to
gauge the influence of individual nodes, we iterate equation (3.5) until $O(n)$ nodes in
the network are reached and use the corresponding values of $m_{\text{new}}$ to rank the nodes.

### 3.4 Branching Process Centrality Versus Other Centrality Measures

Before we present our findings, we dedicate this section to a discussion on the methods we need in order to obtain our results. We discuss the models that we use to simulate the spreading dynamics on the network. We then give the definitions for two quantities used to compare centrality measures as predictors of the most influential nodes in the network. We provide the details and sources of all the network data used in this thesis. We also briefly discuss how the efficiency of the new centrality measure scales with network size. We finish this section by explaining and proving mathematically the connection between the NB and BP centrality at criticality, which is later shown empirically in our results.

#### 3.4.1 The Spreading Dynamics

In order to check the accuracy of each centrality measure, we need to compare the rankings of the nodes given by the centrality measure with the actual list of node rankings from the spreading simulations. We consider the SIR dynamics taking place on the network. As previously discussed in Chapter 1, for this model nodes can be in one of three states - susceptible, infected or recovered. We initialise our system so that we start with only one infected node $i$, and all other nodes are in the susceptible state. After some time, the system reaches a steady state and all the infected nodes eventually recover and the epidemic ends. The spreading power of node $i$ is considered to be the number of nodes whose final state is recovered after running the dynamics. This is repeated for every node in the network and the nodes are then ranked from most influential to least influential depending on the extent of their spreading power.

There are many ways one can simulate SIR dynamics on a computer. We implement three different SIR models in order to simulate the spreading process. For a discrete-time approach, we simulate the SIR dynamics using the ICM (Section 1.4.1). For a continuous-time approach, we simulate the process using the Gillespie Algorithm (Section 1.4.2). As discussed in Chapter 1, the ICM and Gillespie Algorithm simulate different versions of SIR dynamics. The ICM is a synchronous updating model while the Gillespie is an asynchronous updating model. Both models simulate SIR dynamics but in different ways and so are not directly comparable. However, we compare our centrality measure predictions with the list of top ranked nodes from both of these models with input parameters obeying equation (1.51). If both models are simulating the SIR model with comparable input parameters, and provide outcomes that are in agreement, we can use this to support our conjectures regarding the centrality measures.

The third and final model we use to simulate SIR dynamics is based on percolation cluster sizes. It is well known that there is a clear mapping between the steady state properties of the SIR model and percolation. We can use this model as we are consid-
ering the influence of nodes in the long time limit here. The model ranks the nodes according to the size of the cluster to which they belong. More precisely, over many simulations, the most influential node is considered to be the node which belongs to the largest cluster on average. The second most influential node is the node that belongs to the second largest cluster on average and in a similar way all nodes are ranked from most influential to least influential.

We compare the predictions given by the various centrality measures with the actual list of ranked nodes from these three variants of SIR spreading processes. We introduce two quantities used for this comparison in the next section.

### 3.4.2 The Imprecision Function and The Jaccard Distance

Here, we define two quantities. We use them to compare the rankings given by the various centrality measures and the actual ranking of the top-spreaders in the network. The imprecision function and the Jaccard distance produce results based on different criterion. However, they both return a numeric value between 0 and 1 depending on the accuracy of the centrality measures ranking. We specifically decide to use these measures as they are used by Radicchi and Castellano in [51], and we believe our research to be an extension of this work. However, we are aware of the existence of other possible measures that we could use such as the Kendall Tau [60] for example, which measures the correlations between node rankings.

#### 3.4.2.1 The Imprecision Function

The imprecision function $\epsilon(N)$ [51] quantifies the difference between the average size of spreading power initiated by the first $N$ nodes according to a given centrality measure and the average size of the spread from the actual $N$ most influential spreaders from the SIR simulations.

We define $\Upsilon^{(\text{eff})}(N)$ to be the set of actual top $N$ influential spreaders according to the SIR simulations and $\Upsilon^{(x)}(N)$ to be the set of top $N$ influential spreaders according to the centrality measure $x$. We define the imprecision function to be given by

$$
\epsilon^{(x)}(N) = 1 - \frac{Z^{(x)}(N)}{Z^{(\text{eff})}(N)},
$$

(3.9)

where

$$
Z^{(x)}(N) = \frac{1}{N} \sum_{i \in \Upsilon^{(x)}(N)} \langle M_i \rangle,
$$

and

$$
Z^{(\text{eff})}(N) = \frac{1}{N} \sum_{i \in \Upsilon^{(\text{eff})}(N)} \langle M_i \rangle.
$$

Here $M_i$ is the size of the outbreak from node $i$ in both cases. If the centrality $x$ perfectly
predicts the top influential spreaders, the imprecision function equals to zero. The larger
the value for \( \epsilon(x)(N) \), the poorer the the centrality measure \( x \) is for the prediction of top
spreaders.

### 3.4.2.2 The Jaccard Distance

The Jaccard distance \( d_J \) [51] measures the diversity between two sets of nodes \( \Upsilon^{(eff)}(N) \)
and \( \Upsilon^{(x)}(N) \), where \( \Upsilon^{(eff)}(N) \) is the set of actual top \( N \) influential spreaders according
to the SIR simulations and \( \Upsilon^{(x)}(N) \) is the set of top \( N \) influential spreaders according
to the centrality measure \( x \) as before.

The Jaccard distance is given by

\[
d_J(N) = 1 - \frac{\left| \Upsilon^{(x)}(N) \cap \Upsilon^{(eff)}(N) \right|}{\left| \Upsilon^{(x)}(N) \cup \Upsilon^{(eff)}(N) \right|}.
\]

Here \( |\cdot| \) stands for the number of elements in the set. If the two sets are the same,
the Jaccard distance equals to zero. The larger the value of \( d_J(N) \), the poorer the the
centrality measure \( x \) is for the prediction of top spreaders.

### 3.4.3 Network Data

In this section, we clearly present the network data used in this thesis. We provide
detailed descriptions of all the networks here. The reader can refer to this section for
clarification when we present results on some of the below networks later on.

**Synthetic Network 1**

This is a synthetic undirected network. The network is constructed using the configuration
model with a power-law degree distribution, \( P(k) = k^{-\gamma} \), with exponent \( \gamma = 3.5 \).
There are 1,000 nodes in this network.

<table>
<thead>
<tr>
<th>Synthetic Network 1 Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
</tr>
<tr>
<td>1,000</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Undirected</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
<tr>
<td>Unweighted</td>
</tr>
</tbody>
</table>

**Synthetic Network 2**

This is a synthetic undirected network. The network is constructed using the configuration
model with a power-law degree distribution, \( P(k) = k^{-\gamma} \), with exponent \( \gamma = 3.5 \).
There are 2,000 nodes in this network.

<table>
<thead>
<tr>
<th>Synthetic Network 2 Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Nodes</td>
</tr>
<tr>
<td>2,000</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Undirected</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
<tr>
<td>Unweighted</td>
</tr>
</tbody>
</table>
Synthetic Network 3
This is a synthetic Erdös-Rényi random graph. There are 1,000 nodes in this network. The probability that any two links are connected is $p = 0.01$.

<table>
<thead>
<tr>
<th>Synthetic Network 3 Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Nodes</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
</tbody>
</table>

Karate Network
This is the well-known Zachary karate club network. This is an undirected network with 34 nodes. Each node in the network represents a member of the karate club. An edge in the network indicates a tie between two members of the club.

<table>
<thead>
<tr>
<th>Karate Network Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Type</td>
</tr>
<tr>
<td>Edge Type</td>
</tr>
<tr>
<td>Number of Nodes</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
<tr>
<td>Source</td>
</tr>
</tbody>
</table>

Jazz Network
This is an undirected collaboration network between jazz musicians. A node in the network represents a jazz musician. The edges in the network indicate that two musicians played in the same band.

<table>
<thead>
<tr>
<th>Jazz Network Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Type</td>
</tr>
<tr>
<td>Edge Type</td>
</tr>
<tr>
<td>Number of Nodes</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
<tr>
<td>Source</td>
</tr>
</tbody>
</table>

Dolphin Network
This is an undirected social network. The network represents the social interactions of bottlenose dolphins observed between 1994 and 2001. The nodes in the network are dolphins living in Doubtful Sound, a fjord in New Zealand. An edge in the network represents frequent association among dolphins.
Dolphin Network Information

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Dolphins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge Type</td>
<td>Interactions</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>62</td>
</tr>
<tr>
<td>Format</td>
<td>Undirected</td>
</tr>
<tr>
<td>Edges Weights</td>
<td>Unweighted</td>
</tr>
<tr>
<td>Source</td>
<td>[63]</td>
</tr>
</tbody>
</table>

Caltech36 Facebook Network

This is an undirected social network network with 762 nodes. The network represents the Facebook friendship network at California Institute of Technology. The nodes in the network represent Facebook users. The edges represent friendship ties between the Facebook users.

Caltech36 Facebook Network Information

<table>
<thead>
<tr>
<th>Node Type</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge Type</td>
<td>Friendship</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>762</td>
</tr>
<tr>
<td>Format</td>
<td>Undirected</td>
</tr>
<tr>
<td>Edge Weights</td>
<td>Unweighted</td>
</tr>
<tr>
<td>Source</td>
<td>[63]</td>
</tr>
</tbody>
</table>

International E-Road Network

This is an undirected network of 1,177 nodes. The network represents the International E-road numbering system of roads in Europe developed by the United Nations Economic Commission for Europe. The nodes in the network represent the roads. The edges in the network represent connections between the roads.

International E-Road Network Information

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Road</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge Type</td>
<td>Physical Connection</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>1,177</td>
</tr>
<tr>
<td>Format</td>
<td>Undirected</td>
</tr>
<tr>
<td>Edges Weights</td>
<td>Unweighted</td>
</tr>
<tr>
<td>Source</td>
<td>[64]</td>
</tr>
</tbody>
</table>

US Power Grid Network

This an undirected network representing the topology of the Western States Power Grid of the United States. There are 4,941 nodes in the network representing the various substations and the edges represent the high voltage transmission lines that connect them.
US Power Grid Network Information

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Substations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge Type</td>
<td>Transmission Lines</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>4,941</td>
</tr>
<tr>
<td>Format</td>
<td>Undirected</td>
</tr>
<tr>
<td>Edges Weights</td>
<td>Unweighted</td>
</tr>
<tr>
<td>Source</td>
<td>[65]</td>
</tr>
</tbody>
</table>

Face-to-Face Contact Network: 28/04/2009

This network is created using the first day of data collected during the Infectious SocioPatterns experiment which took place in the Science Gallery in Dublin, Ireland in 2009.

Each file in the data set contains a list representing contacts of attendees of the event that lasted at least for a 20 seconds interval on that day. Each line in the file has a time during which the contact was active, a node $i$ and a node $j$, where $i$ and $j$ are represented by node IDs of the people in contact. We take one day set of contacts and then consider the aggregate network where the weight between node $i$ and node $j$ is proportional to the time they spent face-to-face with each other during that day.

<table>
<thead>
<tr>
<th>Face-to-Face Contact Network Information: 28/04/2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Type</td>
</tr>
<tr>
<td>Edge Type</td>
</tr>
<tr>
<td>Number of Nodes</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
<tr>
<td>Source</td>
</tr>
</tbody>
</table>

Face-to-Face Contact Network: 17/07/2009

This network is created using the last day of data collected during the Infectious SocioPatterns experiment.

<table>
<thead>
<tr>
<th>Face-to-Face Contact Network Information: 17/07/2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Type</td>
</tr>
<tr>
<td>Edge Type</td>
</tr>
<tr>
<td>Number of Nodes</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
<tr>
<td>Source</td>
</tr>
</tbody>
</table>
Physician Network

This is a directed network of 246 physicians. The network data was collected to study the innovation spread among physicians in the towns Illinois, Peoria, Bloomington, Quincy and Galesburg. The data was collected in 1966. A node represents a physician. The data set contains an edge-list, there is an edge between two physicians if the left physician said that the right physician was his friend or that he would ask her for advice regarding a topic.

<table>
<thead>
<tr>
<th>Physician Network Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Type</td>
</tr>
<tr>
<td>Edge Type</td>
</tr>
<tr>
<td>Number of Nodes</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
<tr>
<td>Source</td>
</tr>
</tbody>
</table>

Wikipedia Network

Wikipedia is a free encyclopaedia written collectively by volunteers around the world. A small sample of Wikipedia patrons are administrators. These are special users that are granted with access to additional technical features that aid the maintenance of pages. A public vote decides which users can be administrators. The network contains all the Wikipedia voting data from the beginning of Wikipedia until January 2008. Nodes in the network represent Wikipedia users. The directed edges in the network from node $i$ to node $j$ indicates that node $i$ voted for user $j$.

<table>
<thead>
<tr>
<th>Wikipedia Network Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Type</td>
</tr>
<tr>
<td>Edge Type</td>
</tr>
<tr>
<td>Number of Nodes</td>
</tr>
<tr>
<td>Format</td>
</tr>
<tr>
<td>Edges Weights</td>
</tr>
<tr>
<td>Source</td>
</tr>
</tbody>
</table>

Twitter Network

This is a directed network that contains Twitter user-user following information. In this network a node represents a user on Twitter. In the edge-list provided in the data set, an edge indicates that the user represented by the left node follows the user represented by the right node.
3.4.4 Computational Efficiency of the Branching Process Centrality Measure

Many centrality measures have been proposed to date. Centrality measures are used to identify the most influential nodes in a network, with some being better predictors than others, depending on one's definition of influential. However, regardless of their accuracy they need to be easily computed. The computational efficiency of some measures are not adequate when we scale up and deal with large networks. For example, it is well-known that the accuracy of both closeness centrality and betweenness centrality is compromised on large scale networks due to their computational complexity.

Social networks that are representatives of real-world online social media platforms can be very large. As we explain in more detail later on, we believe BP centrality could find a place in this space and be used to accurately identify influential nodes in an online social network setting. Due to the scale of such networks, it is of interest to see how the computational efficiency of the measure scales with network size.

Here, we demonstrate the time efficiency of the BP centrality measure with two separate examples. Firstly, we look at 3 different types of networks constructed using the configuration model with power-law degree distributions of $P(k) \sim k^{-2.5}$, $P(k) \sim k^{-3.5}$ and $P(k) \sim k^{-4.5}$. Within these 3 different configurations, we consider 11 networks of different sizes. We want to test the computational efficiency of the measure as we increase the network size. We begin with a network of 1,000 nodes and increase the network size to 10,000 and from there increase by an increment of 10,000 until we reach a network size of 100,000 nodes. As networks with the same size in different groups have different number of edges (because of the different degree distributions), we look at the number of edges in each network as opposed to the number of nodes for the comparison.

We plot on a logarithmic scale the length of time in seconds it takes to compute the BP centrality versus the number of edges in each network. Here, we label the axes with the time in seconds and the number of edges to make the plot easier to interpret. We can see from Figure 3.5, that there seems to be linear relationship between the logarithm of the time taken to compute the BP centrality on a network and the logarithm of the number of edges in the network in all 3 cases. We also plot the logarithm of the lines $y \propto x$ and $y \propto x^2$ and compare the slopes of these lines with the slopes of our green, pink and yellow lines. We can conclude from this that the relationship between the time and the number edges in the network follows a power-law in which the exponent is less than 2 in all cases. From this example, we also point out the actual computational time. We can see from the plot below that with a network with the order of $1 \times 10^6$ edges,
for the above model parameters, the BP centrality can be computed in an order of minutes.

Figure 3.5: A plot on a logarithmic scale of the time (in seconds) it takes to compute the BP centrality \((\phi = 1)\) for a network versus the number of edges in the network. There are 11 networks dealt with here in total per each group, ranging in sizes of \(n = [1 \times 10^3, 1 \times 10^5]\).

For another example, we also consider 3 different groups of similarly constructed networks. We use a 3-regular graph and a modified 3-regular graph. A 3-regular graph is a network in which each node has exactly three random neighbours. We slightly modify the 3-regular graph to have some diversity in the network. The three networks models we consider are the 3-regular graph and the modified 3-regular graph with every 5 in 2000 nodes being a degree-20 hub node and also every 10 in 2000 nodes being a degree-20 hub node. Again, we look at 11 networks of different sizes within the 3 groups. We deal with the same sizes as before, with the number of nodes in the range \(n = [1 \times 10^3, 1 \times 10^5]\). We plot on a logarithmic scale the time in seconds for the BP centrality to be calculated versus the number of edges in the network.

Again, we can see in Figure 3.6 below that there appears to be a roughly linear relationship between the logarithm of the calculation time and the logarithm of the number of edges in the network. Again, we label the axes with the time in seconds and the number of edges. We plot the logarithm of the lines \(y \propto x\) and \(y \propto x^2\) and compare the slopes of these lines with the slopes of the blue, purple and orange lines below. It appears that the relationship between the time and the number of edges follows a power-law with an exponent less than 2 again here. We also point out that the time taken to calculate the BP centrality on a network with more than \(1 \times 10^6\) edges is only in the order of minutes, for all 3 models considered here.
As discussed in Section 3.3, the BP centrality can also be calculated by solving the linear system in equation (3.8) as opposed to using the iterative method that we have considered. However, calculating the BP centrality by solving the linear system is not feasible for large networks. The calculations to find the inverse of the matrix $I - \phi M$ and also to even store the matrices is not possible even for networks with a size of 100,000 nodes. However, we still want to compare both methods. For that reason, we consider 11 networks, constructed using the configuration model with a degree distribution of $P(k) \sim k^{-3.5}$, in the size range of $n = [1 \times 10^3, 1 \times 10^5]$, and compare the computational efficiency for the calculation of the BP centrality by solving the linear system and using the iterative approach. As we can see in Figure 3.5 below, the time taken to calculate the BP centrality using the iterative approach for networks of this size is negligible. These times could even just be considered the time needed to start and finish the programme. On the other hand, the time taken to calculate the BP centrality by solving the linear system is significant. We were unable to do such a comparison for larger networks as the storage and calculations required to solve the linear system with such large matrices is not computationally feasible. However, we get the idea from Figure 3.7. It shows the drastic difference in computational efficiency when using the iterative method and when solving the linear system.
Figure 3.7: A plot on a logarithmic scale of the time (in seconds) it takes to compute the BP centrality for 11 networks constructed using the configuration model versus the number of edges in the network. The networks here range in sizes of \( n = [1 \times 10^2, 1 \times 10^3] \).

We dedicated this section to the performance of the BP centrality measure as we scale up to large networks. We can see from the examples above in Figure 3.5 and 3.6 that there doesn’t appear to be an exponential relationship between the computational time and the number of edges in a network. The relationship, in fact appears to follow a power-law with an exponent less than 2 in all the cases we considered here. This is a promising result. Not only that but the actual time taken to compute the BP centrality with networks as large as 1,000,000 edges is only in the order of seconds. This makes us confident that the BP centrality (calculated using iteration), can be used on large scale networks without being compromised due to any possible computational complexity.

### 3.4.5 The Non-Backtracking and Branching Process Centrality at Criticality

We dedicate this section to explain and prove some of the findings discussed in [49] and [51]. We then further expand this work and show that the NB and BP centrality coincide at criticality.

Firstly, we show that the NB centrality of a node \( i \) is equal to the probability that node \( i \) is in the percolating cluster at criticality for tree-like networks. We then show that the relative size of the epidemic outbreak starting from a random node \( i \) is equal the NB centrality at criticality on tree-like networks.

Following closely the mathematics in [49], we define \( \pi_i(s) \) to be the probability that node \( i \) belongs to a small percolation cluster of exactly \( s \) nodes.

If the network is a perfect tree (with no loops), then the size \( s \) of the cluster is equal to 1 (for node \( i \) itself) plus the sum of the numbers of nodes reachable along each edge
attached to \( i \), which is zero if the edge is unoccupied or non-zero otherwise. However, if there are loops in the network then this calculation will not, in general, give the exact value of \( s \). In this case, it may be possible to reach the same node along two different occupied edges, which leads to over counting.

For the sake of the following calculations, we propose that the network we are dealing with is a perfect tree. Thus, the probability that node \( i \) belongs to a small percolating cluster of exactly \( s \) nodes, \( \pi_i(s) \), can be written as [49]

\[
\pi_i(s) = \sum_{\{s_j \mid j \in N(i)\}} \prod_{j \in N(i)} \pi_{i \to j}(s_j) \delta\left(s - 1, \sum_{j \in N(i)} s_j\right),
\]

(3.11)

where \( \pi_{i \to j}(s) \) is the probability that exactly \( s \) nodes are reachable along the edge connecting \( i \) and \( j \), and \( N(i) \) is the set of neighbours of node \( i \), and the sum is over all possible sets \( \{s_j \mid j \in N(i)\} \). The delta function guarantees that only sets composed of \( s_j \) values whose sum amounts to \( s - 1 \) have a nonzero contribution.

The probability generating function for \( \pi_i(s) \) is given by

\[
G_i(z) = \sum_{s=1}^{\infty} \pi_i(s) z^s,
\]

(3.12)

\[
G_i(z) = \sum_{s=1}^{\infty} z^s \sum_{\{s_j \mid j \in N(i)\}} \left[ \prod_{j \in N(i)} \pi_{i \to j}(s_j) \right] \delta\left(s - 1, \sum_{j \in N(i)} s_j\right),
\]

(3.13)

\[
G_i(z) = z \prod_{j \in N(i)} \sum_{s_j=0}^{\infty} \pi_{i \to j}(s_j) z^{s_j},
\]

(3.14)

\[
G_i(z) = z \prod_{j \in N(i)} H_{i \to j}(z),
\]

(3.15)

where the probability generating function for \( \pi_{i \to j}(s) \) is given by

\[
H_{i \to j}(z) = \sum_{s=0}^{\infty} \pi_{i \to j}(s) z^s.
\]

(3.16)

We note that no nodes are reachable along \( i \to j \) if the edge between \( i \) and \( j \) is unoccupied (which happens with probability \( 1 - \phi \)) and nonzero otherwise (probability \( \phi \)). This means that \( \pi_{i \to j}(0) = 1 - \phi \), and for \( s > 0 \)

\[
\pi_{i \to j}(s) = \phi \sum_{\{s_k \mid k \in N(j) \setminus i\}} \left[ \prod_{k \in N(j) \setminus i} \pi_{j \to k}(s_k) \right] \delta\left(s - 1, \sum_{k \in N(j) \setminus i} s_k\right).
\]

(3.17)
If we substitute this expression into (3.16) we get [49]

\[ H_{i\to j}(z) = 1 - \phi + \phi z \prod_{k \in N(j) \setminus i} H_{j\to k}(z). \] (3.18)

The equation above suggests a branching process algorithm. We note that the probability that node \(i\) belongs to any small cluster is given by

\[ \sum_{s} \pi_{i}(s) = G_{i}(1) = \prod_{j \in N(i)} H_{i\to j}(1). \]

We can write equation (3.18) as

\[ H_{i\to j}(1) = 1 - \phi + \phi \prod_{k \in N(j) \setminus i} H_{j\to k}(1). \] (3.19)

We know that \(\prod_{j \in N(i)} H_{i\to j}(1)\) is the probability that node \(i\) does not belong to the percolating cluster. We also can see that \(H_{i\to j}(1) = 1\), for all \(i\) and \(j\), is a solution to (3.19), which corresponds the fact that no node is in the percolating cluster.

If the fixed point \(H_{i\to j}(1) = 1\) is stable, then iteration will converge to it and the branching process algorithm tells us that there is no percolating cluster. If it is unstable, we will end up at a different solution and there is a percolating cluster. Thus, the point where \(H_{i\to j}(1) = 1\) goes from being stable to unstable is the percolating threshold.

We can determine the stability of the fixed point by linearising, and write

\[ H_{i\to j}(1) = 1 - \epsilon_{i\to j}. \] (3.20)

We expand equation (3.19) to leading order in \(\epsilon_{i\to j}\) to get [49]

\[ \epsilon_{i\to j} = \phi \sum_{k \in N(j) \setminus \{i\}} \epsilon_{j\to k}. \] (3.21)

In matrix notation this is

\[ \epsilon = \phi \mathbf{M} \epsilon , \] (3.22)

where \(\epsilon\) is the \(2E\)-element vector with elements \(\epsilon_{i\to j}\) and \(\mathbf{M}\) is the \(2E \times 2E\) NB matrix (equation (2.7)). The vector \(\epsilon\) tends to zero and hence the fixed point is stable under iteration if and only if \(\phi\) times the leading eigenvector of \(\mathbf{M}\) is less than one

\[ \frac{\phi}{\lambda_{1}} < 1. \] (3.23)
Therefore, the critical percolation probability $\phi_c$ of a locally tree-like network is equal to the reciprocal of the leading eigenvalue of the NB matrix

$$\phi_c = \frac{1}{\lambda_1}. \quad (3.24)$$

Recall, the probability that node $i$ does not belong to the percolating cluster is given by

$$\prod_{j \in N(i)} H_{i \rightarrow j}(1). \quad (3.25)$$

Therefore, the probability that node $i$ does not belong to the percolating cluster at the threshold can be written as

$$\prod_{j \in N(i)} H_{i \rightarrow j}(1) = \sum_{j \in N(i)} 1 - \epsilon_{i \rightarrow j} = 1 - \sum_{j \in N(i)} \epsilon_{i \rightarrow j}. \quad (3.26)$$

So, the probability that node $i$ does belong to the percolating cluster at the percolating threshold is given by

$$\sum_{j \in N(i)} \epsilon_{i \rightarrow j}. \quad (3.27)$$

The probability that a node $i$ belongs to the percolating cluster at the percolating threshold can also be written as

$$x_i = \sum_j A_{ij} \epsilon_{i \rightarrow j}, \quad (3.28)$$

where $A_{ij}$ are elements of the adjacency matrix and $\epsilon_{i \rightarrow j}$ are the elements of the leading eigenvector of the NB matrix. So at criticality the probability that a node $i$ is in the percolating cluster at the threshold is the NB centrality of node $i$ as previously defined in equation (2.12) in Section 2.3.2.

If the giant cluster in a percolation process occupies a fraction $\langle Q_i \rangle$ of the entire network then a randomly chosen initial node $i$ will fall within it with probability $\langle Q_i \rangle$. And if it does then the disease will spread to infect the entire giant cluster, creating an epidemic reaching a fraction of the population also equal to $\langle Q_i \rangle$.

Therefore, the relative size of an epidemic outbreak starting from an initial node $i$ at the threshold is equal to the probability that node $i$ belongs to the percolating cluster. This coincides with the NB centrality at the critical point. Thus
As a consequence the top spreaders are the nodes with the highest NB centrality. Note that this is an exact result, provided the network structure is locally tree-like.

We now examine the mathematics used to calculate the BP centrality, and compare both the NB and BP centrality at criticality. In order to determine the influence of a node \( i \) for the BP centrality measure at criticality, we calculate \( m_{i \rightarrow j}(a) \), the expected size of a tree (infection) created by a transmission attempt from \( i \) to \( j \) at age \( a \). Summing over all \( m_{i \rightarrow j}(a) \), where \( j \) denotes the nearest neighbours of node \( i \), and including \( i \) itself, we can calculate the tree size created from seed \( i \) at age \( a \).

Recalling equation (3.3) from Section 3.2, we can write an equation for \( m_{i \rightarrow j}(a) \) at criticality as

\[
m_{i \rightarrow j}(a) = \phi_c \left( 1 + \sum_{k \in N(j) \setminus \{i\}} m_{j \rightarrow k}(a-1) \right).
\] (3.30)

Considering the steady state of the system as we did for the NB centrality, we get the equation

\[
m_{i \rightarrow j}(\infty) = \phi_c \left( 1 + \sum_{k \in N(j) \setminus \{i\}} m_{j \rightarrow k}(\infty) \right).
\] (3.31)

Therefore, the expected tree size created from the seed node \( i \) in the steady state (which is equivalent to the size of the percolation cluster starting from an initial node \( i \) as previously discussed), is therefore given by

\[
S_i = \sum_{j \in N(i)} m_{i \rightarrow j}(\infty) + 1.
\] (3.32)

The expected tree size from node \( i \) is equal to node \( i \) itself plus the sum over all expected tree sizes that were a result of a transmission from node \( i \) to all of its nearest neighbours. We define \( S_i \) to be the BP centrality of node \( i \). Therefore as a consequence the top-spreaders are the nodes with the highest BP centrality.

At criticality, \( x_i \) which is the NB centrality is the relative size of an epidemic outbreak starting from an initial node \( i \), and \( S_i \) which is the BP centrality is the expected tree size created from an initial transmission from node \( i \). We arrive independently at an estimation for the outbreak size in calculations for the NB and BP centrality at criticality. Consequently, both measures should give the same node rankings in this case.

The above work is a qualitative argument that shows how the NB and BP coincide at criticality. We now attempt to show this more rigorously by means of a mathematical
proof.

Recalling equation (3.8) in Section 3.3, which is

$$(I - \phi M) m_\infty = \phi I.$$  

(3.33)

This is the linear system used to calculate the BP centrality.

Let us assume $M$ has $2E$ orthogonal eigenvectors that form a basis for $\mathbb{R}^{2E}$. Then we can express any vector in $\mathbb{R}^{2E}$, for example $\phi I$, as a combination of the basis set $\{v_1, v_1, \cdots, v_{2E}\}$. We can therefore write

$$\phi I = \sum_{j=1}^{2E} d_j v_j, \text{ for some coefficients } d_j.$$  

(3.34)

Similarly, we write

$$m_\infty = \sum_{j=1}^{2E} c_j v_j, \text{ for some coefficients } c_j.$$  

(3.35)

Substituting this into equation (3.33), we get

$$(I - \phi M) m_\infty = \sum_{j=1}^{2E} d_j v_j$$

$$\Rightarrow (I - \phi M) \sum_{j=1}^{2E} c_j v_j = \sum_{j=1}^{2E} d_j v_j$$

$$\Rightarrow \sum_{j} c_j (v_j - \phi M v_j) = \sum_{j=1}^{2E} d_j v_j$$

$$\Rightarrow \sum_{j} c_j (1 - \phi \lambda_j) v_j = \sum_{j=1}^{2E} d_j v_j$$  

(3.36)

By the orthogonality assumption, we therefore must have

$$c_j = \frac{d_j}{1 - \phi \lambda_j}, \text{ for all } j = 1, 2, \ldots, 2E.$$  

(3.37)

As $\phi$ approaches $\frac{1}{\lambda_1}$ from below, we note that $c_1 \to \infty$, while all other $c_j$ ($j \neq 1$) remain
finite. Thus, \( m_\infty = \sum_j c_j v_j \) becomes parallel to \( v_1 \), since the sum is dominated by the \( c_1 \) coefficient.

Hence, at criticality (\( \phi = \frac{1}{\lambda_1} \)), we find \( m_\infty \) is proportional to \( v_1 \). Therefore, the BP centrality is proportional to \( v_1 \) at criticality. We also know from equation (3.28) above that the NB centrality is also proportional to largest leading eigenvector of the NB matrix, \( M \). And so the rankings given by the BP and NB centralities at criticality are equal.

However, note that our assumption on the eigenvectors forming a basis is not fully tested. It would be true if \( M \) were a symmetric matrix, but in general the non-backtracking matrix is not symmetric, so this assumption requires further research (but this is beyond the scope of this thesis).

In this section, we investigated the NB and BP centrality at criticality. Using a qualitative argument, we show how the NB and BP centrality coincide at criticality. We then went further and give a grounded mathematical condition for when the centralities gave equal rankings. We see later that the NB and BP centrality do coincide on various networks in the results we present later (Section 3.5.0.3).

### 3.5 Results: Performance of the Branching Process Centrality

In this section, we present our results. We compare the performance of our proposed centrality measure based on a discrete branching process model with various existing centrality measures (Chapter 2) as predictors of the top spreaders in the network.

We present the results obtained when using the imprecision function and Jaccard distance for comparing the predictions given by various centrality measures. We simulate three variations of the SIR model with comparable parameters on the same network for an extensive comparison. We simulate each SIR model for 1,000 realisations. We then, plot the imprecision function and Jaccard distance for the various centralities as a function of \( N \). Firstly, we show results on some synthetic networks. In the subsequent section, we present results on various real-world networks.

#### 3.5.1 Results: Synthetic Networks

We construct an undirected network of 1,000 nodes using the configuration model. We set the degree distribution to follow a power-law degree distribution \( P(k) = k^{-\gamma} \), with exponent \( \gamma = 3.5 \). We call this network the Synthetic Network 1. A graph of the network can be seen below in Figure 3.8.
Figure 3.8: Synthetic Network 1 - An undirected network of 1,000 nodes. This network is constructed using the configuration model with a power-law degree distribution.

As we have seen in Chapter 2, there exists an abundance of centrality measures and so it would not be possible to compare and contrast them all. We deal with six centrality measures including our newly proposed centrality measure based on a branching process approach. We choose these specific centralities as they cover measures based on network structure, location of nodes with respect to others and eigenvector based centralities. We consider degree, closeness, betweenness, eigenvector, NB and our new BP centrality.

Figure 3.9: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Synthetic Network 1, of 1,000 nodes for the three SIR spreading models - Gillespie ($\beta = 0.1$ and $\gamma = 1$), ICM and Percolation ($\phi = 0.1$).
We plot the imprecision function and the Jaccard distance for each of these centralities as a function of the number of top influential seeds ($N$). Specifically, we consider $N$ in the range of 1 to 100 here. So, we compare the top 1 to 100 influentially ranked nodes from the SIR simulations with the predictions from the corresponding centrality measures.

A centrality measure which is a good predictor of the most efficient spreaders in a network will result in an imprecision function and Jaccard distance value close to zero. We can clearly see in Figure 3.9, for the imprecision function, the BP centrality performs the best compared to all other centrality measures considered on all three SIR simulation models. The light blue line is closest to zero for all values of $N$. Also, examining Figure 3.10 for the Jaccard distance, the BP centrality again is superior on all three SIR models. The results from both measures are in agreement. For the specified model parameters, the BP centrality is the best predictor of the top influential spreaders in all cases according to the values of the imprecision function and the Jaccard distance.

We also consider an undirected network of 2,000 nodes constructed using the configuration model. As before, we set the degree distribution to follow a power-law degree distribution $P(k) \sim k^{-\gamma}$, with exponent $\gamma = 3.5$. We call this network the Synthetic Network 2. A graph of the network can be seen in Figure 3.11.

We deal with the same centrality measures as we did before. We plot the imprecision function and Jaccard distance for each centrality as a function of $N$, with $N$ ranging from 1 to 100.
Figure 3.11: Synthetic Network 2 - An undirected network of 2,000 nodes. This network is constructed using the configuration model with a power-law degree distribution.

Figure 3.12: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Synthetic Network 2, of 2,000 nodes for the three spreading models - Gillespie ($\beta = 0.1$ and $\gamma = 1$), ICM and Percolation ($\phi = 0.1$).
Figure 3.13: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the Synthetic Network 2, of 2,000 nodes for the three spreading models - Gillespie ($\beta = 0.1$ and $\gamma = 1$), ICM and Percolation ($\phi = 0.1$).

We can see in Figure 3.12, for the imprecision function, the BP centrality performs the best compared to all other centrality measure considered on all three SIR simulation models. Also, examining Figure 3.13, for the Jaccard distance, the BP centrality again is generally superior on all three SIR models.

We look at an undirected network of 1,000 nodes constructed using the Erdős-Rényi random graph model. In this model, any two nodes in the network are connected with a constant probability $p$. In this case we set $p=0.01$. We call this network the Synthetic Network 3. Figure 3.14 shows a plot of the random graph.

Figure 3.14: Synthetic Network 3 - An undirected network of 1,000 nodes. This network is constructed using the Erdős-Rényi random graph model with $p=0.01$. 
We consider the same centralities as in the previous examples. Again, we plot the imprecision function and Jaccard distance for each centrality as a function of $N$, with $N$ ranging from 1 to 100.

![Imprecision Function Gillespie](image1)

![Imprecision Function ICM](image2)

![Imprecision Function Percolation](image3)

**Figure 3.15**: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Synthetic Network 3, of 1,000 nodes for the three spreading models - Gillespie ($\beta = 0.01$ and $\gamma = 1$), ICM and Percolation ($\phi = 0.01$).

![Jaccard Distance Gillespie](image4)

![Jaccard Distance ICM](image5)

![Jaccard Distance Percolation](image6)

**Figure 3.16**: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the Synthetic Network 3, of 1,000 nodes for the three spreading models - Gillespie ($\beta = 0.01$ and $\gamma = 1$), ICM and Percolation ($\phi = 0.01$).

From Figure 3.15 and Figure 3.16, we can see that the BP centrality performs well, returning low values for the imprecision function and the Jaccard distance. However, the other centrality measures generally also perform well on this network. On all three SIR simulation models, the BP, NB and eigenvector centralities predominately give values...
of the imprecision function and Jaccard distance closest to zero however there are a few values of \( N \) when other centralities are slightly superior.

### 3.5.2 Results: Real World Networks

In the previous section, we compared the performance of the various centrality measures on some synthetic networks. In this section, we examine the predicting power of the centrality measures on real-world networks. We consider a variety of real-world networks.

The first network we look at is a social network of friendships between 34 members of a karate club in the 1970’s [61]. A graph of the network can be seen below in Figure 3.17. Each node in the network is a karate club member and the edges indicate friendships between members.

![Figure 3.17: Karate Club Network.](image)

For consistency, throughout the entire set of results presented in this thesis we consider \( N \) to range from 1 to 100. However, like the above network we will deal with a few examples where there are less than 100 nodes in the network. In these cases, we consider \( N \) to range from 1 to the size of that network. Since this a network of only 34 nodes, we consider values of \( N \) to range from 1 to 34 here.
For the imprecision function, we can see in Figure 3.18, that BP centrality performs generally very well consistently over all three spreading models. Degree, eigenvector and NB centrality also return quite low values and hence act as good predictors for the top ranked nodes. Looking at Figure 3.19, degree, eigenvector, NB and BP centrality give relatively low values for the Jaccard distance.

An important thing to note when analysing these similarity measure plots is the difference between the imprecision function and Jaccard distance. The imprecision function...
quantifies the similarity between the influence of the top ranked nodes given by the centrality measure and the numerical simulations. It is clear here that the BP centrality predicts this influence very well. However, looking at the results given by the Jaccard distance gives us additional information. The Jaccard distance measures how accurately the centrality measures are at actually predicting the top nodes and not just the influence. In this case eigenvector, BP and degree centralities give the best results but the plots are quite noisy with fluctuations away from zero. The reason for the Jaccard distance giving worse results for the centralities is most likely due to degeneracy of influence in the network. There are many nodes with the same spreading power and this leads to slight discrepancies between the list of ranked nodes given by the numerical simulations and the centrality measures. In this case of degeneracy, it may be more insightful to look at the results given by the imprecision function if seeking the top influential nodes. One can then check results given by the Jaccard distance for the accuracy of individual node ranking predictions.

The next network we examine is a collaboration network of popular jazz musicians. The network comprises of 198 bands from 1912-1940 [32]. The nodes in the network are the bands and the edges indicate if there is at least one common musician between bands.

Figure 3.20: Jazz Network.
It is clear from Figure 3.21 and 3.22, for the specified parameters that the BP centrality and also degree centrality are the superior predictors for the most efficient spreaders in the network.

We consider a social network of 62 bottlenose dolphins. The nodes are the bottlenose dolphins of the bottlenose dolphin community living off Doubtful Sound, a fjord in New Zealand. The links in the network represent the associations between dolphins. The
dolphins were observed between 1994 and 2001 [63].

Figure 3.23: Dolphin Network.

Figure 3.24: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Dolphin Network, of 62 nodes for the three spreading models - Gillespie ($\beta = 0.1 \gamma = 0.8$), ICM and Percolation($\phi = 0.1$).
There are 62 nodes in this network so we consequently consider values of $N$ to range from 1 to 62. The BP centrality returns the lowest values for the imprecision function and Jaccard distance overall the three spreading dynamics, albeit others performing as well for certain values of $N$ (Figure 3.24 and Figure 3.25). We can conclude that the BP centrality is the best centrality for the prediction of the most influential dolphins in the network.

Another social network we consider is the social friendship network at Caltech University extracted from Facebook. The network consists of 762 people with edges representing friendship ties [63].
Figure 3.26: Caltech36 Facebook Network.

Figure 3.27: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Caltech36 Facebook Network, of 762 nodes for the three spreading models - Gillespie ($\beta = 0.01, \gamma = 1$), ICM and Percolation ($\phi = 0.01$).
As in the Karate Club network, there are few centralities that predict to a similar accuracy the top spreaders in the network. We can see clearly from Figure 3.27 and Figure 3.28, that the imprecision function and Jaccard distance give low values for BP, NB, degree, and eigenvector centralities. These measures can be considered the most accurate predictors of the top spreaders on the Caltech36 Facebook Network.

Social networks such as Facebook are well known for their high level of structural clustering. The existence of short loops such as triangles are common in social networks. This corresponds to the overlapping of friends between two people linked on Facebook, which is a common occurrence in reality. As previously discussed in Section 3.2, the BP centrality is based on a branching process approach. It assumes that the network is tree-like with the independence of children nodes. The Facebook social network violates this assumption. This is a reason for the poorer performance of the BP centrality on some networks like Facebook. However, another important point to note is that when in the sub-critical regime the chances of spreading through these loops is small. In these instances even social networks with high clustering can be considered tree-like and hence the BP centrality can return accurate results.

The International E-road network is a numbering system of roads in Europe developed by the United Nations Economic Commission for Europe [64]. There are 1,177 nodes in the network representing roads and the edges in the network represent links between the roads.
Figure 3.29: International E-Road Network.

Figure 3.30: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the E-Road Network, of 1,177 nodes for the three spreading models - Gillespie($\beta = 0.25 \gamma = 1$), ICM and Percolation ($\phi = 0.2$).
Figure 3.31: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the E-Road Network, of 1,177 nodes for the three spreading models - Gillespie ($\beta = 0.25 \, \gamma = 1$), ICM and Percolation ($\phi = 0.2$).

From Figure 3.30, we can say that the BP centrality is not the best predictor of the top influence according to the results from the imprecision function. Degree centrality is the superior measure here. Looking at Figure 3.31 for the Jaccard distance, we can see the same here. For both the imprecision function and the Jaccard distance, the degree centrality is the most accurate predictor of the top spreaders. However the BP centrality is also quite an accurate measure according to both quantities, with the other centrality measures returning quite inaccurate results. The lower performance of BP can be due to the very loose connectivity of modules in this network; such a finite size effect was shown to decrease the accuracy of conventional mean-field and message-passing methods that provide a thermodynamic limit description [75].

The last real-world network we consider here is a network representing the topology of the Western States Power Grid of the United States. There are 4,941 nodes in the network representing the various substations and the edges represent the high voltage transmission lines that connect them [65].
Figure 3.32: US Power Grid Network.

Figure 3.33: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the US Power Grid Network, of 4,941 nodes for the three spreading models - Gillespie ($\beta = 0.1$ and $\gamma = 0.8$), ICM and Percolation ($\phi = 0.1$).
There is no ambiguity here. It is clear from Figure 3.33 and 3.34, that the BP centrality is the optimum predictor of top spreaders in the network. The imprecision function and Jaccard distance are the lowest for the specified values of $N$ across all three models for the BP centrality. We can deduce that the BP centrality is the most accurate predictor of the highly important substations on the US power grid network.

### 3.5.3 Results: Sub-Critical, Critical and Super-Critical Regime

In the previous sections, we examined the performance of the various centrality measures for the three SIR model variations on synthetic and real-world networks. We considered arbitrary input parameters (transmission rate $\beta$ and recovery rate $\gamma$) for the Gillespie algorithm and by taking $\phi = \frac{\beta}{\beta + \gamma}$ (Chapter 1) as inputs for the ICM and Percolation simulations.

In this section, we consider the three SIR processes taking place in the sub-critical, critical and super-critical regime on one synthetic network and two real-world networks from previous sections. As discussed in Chapter 2, Karrer et al. in [49] prove that the critical transmission probability $\phi_c$ is given by the inverse of the leading eigenvalue of the NB matrix in tree-like networks (equation 3.24). Note that we have derived an equation for $\phi_c$ in Chapter 1, however this is an approximate result that assumes the children nodes are independent of each other. We use the result posed by Karrer et al. to estimate $\phi_c$ because they show this is an exact result for tree-like networks and is also a very good approximation in networks that aren’t necessarily tree-like. We take $\phi_c = \frac{1}{\lambda_1}$, where $\lambda_1$ is the leading eigenvalue of the NB matrix.

We deal with input parameters $0.5\phi_c$ (sub-critical), $\phi_c$ (critical) and $1.5\phi_c$ (super-critical). By choosing these fractions of the critical probability as inputs, we can compare the results on the various networks under consideration.
As previously discussed and shown in Section 3.4.5, Radicchi and Castellano in [51] prove that the NB centrality is the optimum criterion for the identification of influential spreaders at criticality (equation 3.27). In this section, we see how our newly proposed BP centrality compares to the NB centrality at criticality and also consider its behaviour in the sub-critical and super-critical regimes.

**Synthetic Network 1 - 1,000 nodes with a power-law degree distribution**

We consider the Synthetic Network 1, of 1,000 nodes from Section 3.4.3. We calculate the critical probability by taking the inverse of the leading eigenvalue of the NB matrix of the network. We found the critical probability to be $\phi_c = 0.43$ and also consider $0.5\phi_c = 0.215$ and $1.5\phi_c = 0.645$ as inputs to the SIR simulation models. The figures below show plots for the imprecision function and the Jaccard distance for the three SIR models with input parameters $0.5\phi_c$, $\phi_c$ and $1.5\phi_c$.

![Imprecision Function Gillespie](image)

Figure 3.35: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Synthetic Network 1, of 1,000 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 0.5\phi_c$. 

![Imprecision Function ICM](image)

![Imprecision Function Percolation](image)
Figure 3.36: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the Synthetic Network 1, of 1,000 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 0.5\phi_c$.

Figure 3.37: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Synthetic Network 1, of 1,000 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = \phi_c$. 

85
Figure 3.38: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the Synthetic Network 1, of 1,000 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = \phi_c$.

Figure 3.39: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Synthetic Network 1, of 1,000 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 1.5\phi_c$. 
In the sub-critical regime, the BP centrality is clearly the superior centrality measure across all three SIR processes for the imprecision function and the Jaccard distance (Figure 3.35 and Figure 3.36). At criticality, the NB and BP centrality present identical results. The BP centrality and NB centrality coincide for the imprecision function and Jaccard distance values (Figure 3.37 and Figure 3.38).

In the super-critical regime, the BP centralities accuracy depreciates drastically and doesn’t appear to perform well. As explained in Section 3.2, the BP centrality is based on a branching process approach. It assumes the network is tree-like and hence the absence of any short loops or triangles. In the super-critical regime, the spreading process is vast and the chances of spreading through loops if they exist in the network is high unlike in the sub-critical regime. So in the super-critical regime, if the children nodes in the network are not independent, the assumption of a tree like network is violated and the BP centrality returns inaccurate results (Figure 3.39 and Figure 3.40).

**Dolphin Network**

We consider the Dolphin Network from Section 3.4.3. We calculate the critical transmission probability to be $\phi_c = 0.189$ and also take $0.5\phi_c = 0.0945$, $1.5\phi_c = 0.2835$ as inputs for the SIR simulation models. As before the figures below show plots for the imprecision function and the Jaccard distance for the three SIR models with input parameters $0.5\phi_c$, $\phi_c$ and $1.5\phi_c$. 

---

**Figure 3.40:** A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the Synthetic Network 1, of 1,000 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 1.5\phi_c$. 

---

Dolphin Network from Section 3.4.3. We calculate the critical transmission probability to be $\phi_c = 0.189$ and also take $0.5\phi_c = 0.0945$, $1.5\phi_c = 0.2835$ as inputs for the SIR simulation models. As before the figures below show plots for the imprecision function and the Jaccard distance for the three SIR models with input parameters $0.5\phi_c$, $\phi_c$ and $1.5\phi_c$. 

---

87
Figure 3.41: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Dolphin Network, of 62 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 0.5\phi_c$.

Figure 3.42: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the Dolphin Network, of 62 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 0.5\phi_c$. 

88
Figure 3.43: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Dolphin Network, of 62 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = \phi_c$.

Figure 3.44: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the Dolphin Network, of 62 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = \phi_c$.
Figure 3.45: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the Dolphin Network, of 62 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 1.5\phi_c$.

Figure 3.46: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the Dolphin Network, of 62 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 1.5\phi_c$.

At criticality the NB and BP centrality coincide exactly for both the imprecision function
and the Jaccard distance. As per [51], the NB centrality is proven to be the optimum centrality measure at criticality in tree-like networks. Our newly proposed BP centrality provides the same accuracy, giving the same results as the NB centrality at criticality. Our BP centrality is superior to the NB centrality and all other centralities considered in the sub-critical regime. It seems reasonable to suspect from these results that the BP centrality is a more accurate measure for the identification of influential spreaders in the sub-critical regime. Unlike the other centralities here, our new BP centrality doesn’t just take into account the structure of the network. It considers the dynamics taking place on the network. We propose this to be a possible reason for its superiority. We can see that in the super-critical regime that the performance of BP is surprisingly better than the other centrality measures across all processes for the imprecision function and the Jaccard distance. However, it is important to note that the accuracy of the BP centrality deteriorates in the super-critical regime. The reasons for this are the same as discussed for the Synthetic Network 1 in the super-critical regime.

**US Power Grid Network**

We consider the US Power Grid Network from Section 3.4.3. We calculate the critical probability $\phi_c = 0.161$ and $0.5\phi_c = 0.0805$, $1.5\phi_c = 0.2415$ as inputs for the SIR simulation models. As before the figures below show plots for the imprecision function and the Jaccard distance for the three SIR models with input parameters $0.5\phi_c$, $\phi_c$ and $1.5\phi_c$.

![Figure 3.47: A plot of the imprecision function versus N. The three plots show the results of the various centralities on the US power grid Network, of 4,941 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 0.5\phi_c$.](image-url)
Figure 3.48: A plot of the Jaccard distance versus \( N \). The three plots show the results of the various centralities on the US power grid Network, of 4,941 nodes for the three spreading models - Gillespie, ICM and Percolation with \( \phi = 0.5\phi_c \).

Figure 3.49: A plot of the imprecision function versus \( N \). The three plots show the results of the various centralities on the US power grid Network, of 4,941 nodes for the three spreading models - Gillespie, ICM and Percolation with \( \phi = \phi_c \).
Figure 3.50: A plot of the Jaccard distance versus $N$. The three plots show the results of the various centralities on the US power grid Network, of 4,941 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = \phi_c$.

Figure 3.51: A plot of the imprecision function versus $N$. The three plots show the results of the various centralities on the US power grid Network, of 4,941 nodes for the three spreading models - Gillespie, ICM and Percolation with $\phi = 1.5\phi_c$. 

93
Again, we can confirm what we have seen for the previous networks. The NB and BP centrality coincide here again at criticality. In [51], they prove that NB centrality is the most accurate centrality measure at criticality. We show theoretically in Section 3.4.5 that the BP centrality coincides with the NB centrality at criticality. The results show this here. In fact, as before, our BP centrality is superior to the NB and all other centrality measures considered in the sub-critical regime again for this network. For the super-critical regime, eigenvector, NB and BP are the best performing measures, with the accuracy of the BP centrality getting worse after the critical point. As already discussed, this depreciation is due to the vast spreading through any possible loops in the network and this violates the BP centrality assumptions. However, it is clear that BP centrality is one of the methods in this regime.
Chapter 4

The Performance of the BP Centrality on More Complex Spreading Processes

4.1 Performance of the Branching Process Centrality on Face-to-Face Contact Networks

We now look to test the performance of the BP centrality on real-world networks where the links between individual nodes are weighted, (links between some nodes are stronger than others). We consider a dataset that contains the daily dynamic contact networks collected during an Infectious SocioPatterns experiment. The event took place at the Science Gallery in Dublin, Ireland from the 28th of April 2009 to the 17th of July 2009, during the arts science exhibition INFECTIOUS: STAY AWAY [66].

The dataset contains numerous files, with each file representing the date in which the experiment took place. Each file contains a list representing contacts during 20 second intervals for that day. Each line in the file has a time during which the contact was active, a node \( i \) and a node \( j \), where \( i \) and \( j \) are represented by node IDs of the people in contact.

In terms of our analysis, we take one day set of contacts and then consider the aggregate network where the weight between node \( i \) and node \( j \) is proportional to the time they spent face-to-face with each other during that day.

Using these weights multiplied by a common global factor as probabilities in the ICM, we look to see how the BP centrality behaves compared to the other structurally based centralities (now also taking into account the weights). We only investigate dynamics in the sub-critical regime as this is where the BP centrality seems to outperform the other measures and where we believe it should find its place in future research.

We consider data from the first day (28/04/2009) and the last day (17/07/2009) of the experiment. Here, we consider the ICM to implement the SIR spreading dynamics. We use a small constant global factor \( \phi \), in order to ensure that the system is in the sub-critical regime.
**First Day 28/04/2009**

**Imprecision Function**
\[ \phi = 0.009 \]

**Jaccard Distance**
\[ \phi = 0.009 \]

Figure 4.1: A plot of the imprecision function and Jaccard distance for the first day of the experiment. We use a small constant global factor of \( \phi = 0.009 \).

**Last Day 17/07/2009**

**Imprecision Function**
\[ \phi = 0.005 \]

**Jaccard Distance**
\[ \phi = 0.005 \]

Figure 4.2: A plot of the imprecision function and Jaccard distance for the last day of the experiment. We use a small constant global factor of \( \phi = 0.005 \).

The plots above in Figure 3.53 and Figure 3.54 are graphs of the imprecision function and Jaccard distance versus \( N \), using the aggregate contact networks from the first day and the last day of the experiment respectively. Since the networks are weighted in this
scenario (each link between pairs of nodes is proportional to the time the nodes spent in contact), we consider weighted degree (W Degree), weighted closeness, weighted betweenness and weighted eigenvector centrality and un-weighted degree (UW Degree) for a fair comparison.

Recall, equation (3.3) used to calculate the BP centrality from Section 3.2 is given by

\[ m_{i \rightarrow j}(a) = \phi_{i \rightarrow j} \left( 1 + \sum_{k \in N(j) \setminus \{i\}} m_{j \rightarrow k}(a-1) \right). \]

In order to calculate the size of a tree created by a transmission from node \( i \) to node \( j \), we need to know the transmission probability \( \phi_{i \rightarrow j} \) along that edge. In this case, we take \( \phi_{i \rightarrow j} \) to be the constant global factor multiplied by the weight of the link between node \( i \) and node \( j \), \( (\phi_{i \rightarrow j} = \phi \times \text{(weight of the link } i \rightarrow j)) \). This corresponds to an infected individual being more likely to infect a susceptible neighbour if they spend more time together. This is a plausible assumption, and it is of interest how the BP centrality performs when taking into account this inhomogeneous spreading process. We use the \( a \to \infty \) limit here because we are ranking the nodes in terms of their spreading power in the steady state of the system. We could also consider a time-dependent centrality here, and analyse the influence of nodes at different times of the spreading process. We will discuss this later as a possible avenue for future work.

It is clear from Figure 3.53 that the BP centrality performs the best compared to the other centrality measures for the first day of the experiment. The weighted degree and closeness are the next best performing centralities in this scenario. It is the same case in Figure 3.54, for the last day of the experiment. The BP centrality performs the best compared to the other centrality measures when predicting the most influential nodes in the contact networks of the first day and the last day of the experiment.

The inhomogeneous spreading process dealt with here could be considered a more realistic portrayal of real-world social dynamics. It seems plausible to assume that in the social networks setting, people are more likely to get ‘infected’ with a trend or opinion by people they spend more time with than those they only interact with occasionally. In the cases presented here, the BP centrality performs very well. This motivates the work of the next chapter. There, we examine the performance of the BP centrality on a variety of social networks where the spreading model follows empirically and analytically supported dynamics.

### 4.2 Social Networks: Degree-Dependent Dynamics

The emergence of online social media has greatly enhanced the ability to communicate with one another. The ease of contacting each other allows the immediate and successful spread of information. We can now receive information from multiple resources at once. Having a collection of apps at our fingertips, such as Twitter, Facebook and Instagram for example, means the flow of information is highly competitive with different topics
striving for our attention. This makes the dynamics of information on these platforms complicated. In fact, very little is known about the exact driving factors for the dynamics of information spread in these communication systems. In [67], they present an analytic model for social spreading behaviour. This model reproduces many of the fundamental characteristics of empirical data for the the dynamics of information flow in communication media.

We can think of the spread of information on social networks like a cascade. A person adopts an attitude or a behaviour. The avalanche spreads if the friends of that person also adopt this new attitude and in turn pass on this social influence effect to their own friends, who may further propagate the attitude. As previously discussed in Chapter 1, the description of information spreading is like that of epidemics of disease. In this chapter, we use the disease spread framework to model the spread of information. We modify the disease spread dynamics used up to this point to account for the analytic findings of [67], which describe social spreading phenomena.

In order to account for the dynamics found in social networks, we need to slightly modify the disease spread models. We consider the ICM with non-uniform probabilities of transmission. In [67], they concern themselves with Twitter networks, which are directed networks. They show in equation (22) of [67], that the probability of a given meme (attitude or behaviour) being chosen for retweet is inversely proportional to the in-degree $j$ of the node (at least in the large-$j$ limit, or if the innovation probability $\mu$ is small). Here, we are only dealing with simple spreading dynamics, and so do not concern ourselves with the possibility of attitude innovation along the spreading process. In this case with the absence of any innovation, we can deduce from [67] a possible degree-dependent ‘transmission probability’. Along with the analytic work of [67], there is also empirical evidence of this degree-dependence for directed Twitter networks in [68].

Thus far in our work, we have been focusing our attention only on undirected networks. In [67], they generalise their work of directed networks to undirected networks. If we look at the argument leading to equation (60) in [67], they show that the probability of a given message being chosen is, in this case, inversely proportional to the degree $k$ of the node. For undirected networks, we propose to look at a transmission model where the probability of transmission from node $i$ to node $j$ is inversely proportional to the degree of node $j$.

As well as the intrinsic dynamics, there can be a fundamental global characteristic difference between information spread on social networks and disease spread on social networks. The spread of information on social networks can result in very different global outcomes compared to the spread of an epidemic disease. As we saw in Chapter 1 in the case of epidemics, there either tends to be an epidemic which in turn infects the entire population or at least a large portion of the population or there is no epidemic at all. In the case of information spreading on social networks, the latter is most likely always the case. If we consider the entire social network of Twitter for example, a meme created by a seed individual is not likely to propagate through the entire social network but only a fraction of it. The followers of the seed may or may not retweet the meme to their followers with chance of further propagation. However, often this meme will only
reach a tiny fraction of the Twitter network before dying off.

We have seen thus far that the BP centrality performs the best in the sub-critical regime. As already explained in detail in Chapter 3, this is because the measure is based on a branching process method which assumes a tree-like network structure with children nodes independent of each other. In the sub-critical regime, many networks can be considered tree-like because the probability of successfully spreading through loops in the network is very small. The errors due to the branching process approximation made in the BP centrality therefore become negligible in the sub-critical regime. As explained above, spreading process on social networks tend to be in the sub-critical regime, and so we propose to test the BP centrality on a variety of social networks where the propagation of information is sub-critical.

For the remainder of this chapter, we investigate the performance of the BP centrality on undirected networks in the sub-critical regime where the transmission probability from node $i$ to node $j$ is inversely proportional to the degree of node $j$. We then extend our work to directed networks, and test the performance of the BP centrality on networks in the sub-critical regime where the transmission probability from node $i$ to node $j$ is inversely proportional to the in-degree of node $j$. The aim of the following sections of this chapter is to examine the performance of the BP centrality in the sub-critical regime when the spreading dynamics are degree-dependent. It is of interest to us to see whether the BP centrality is a good identifier of the top spreaders in a social network when the spreading model is modified to display dynamics that are characteristic of real-world social networks.

### 4.3 Results: Undirected Networks with Degree-Dependent Dynamics

In this section, we remain focused on networks that are undirected. To this point, we have examined the performance of the BP centrality on undirected networks where the spreading dynamics follow the basic SIR model. We now look to extend our results and study the performance of the BP centrality on undirected networks where the dynamics are slightly modified.

As discussed above, it has been shown analytically in [67] that for some undirected social networks, the probability of an attitude or behaviour being chosen is inversely proportional to the degree $k$ of the node. We propose a small adjustment to the ICM so that the transmission probability between node $i$ and node $j$ is inversely proportional to the degree $k_j$ of node $j$. In this case, our model follows the analytical finding of [67] which is claimed to be characteristic of a class of real-world dynamics on social networks.

Recalling the equation used to calculate the BP centrality is given by equation (3.4). Using the required transmission probability, the equation we use to calculate the BP centrality here is given by
\[
m_{i\rightarrow j}(a) = \frac{\phi}{k_j} \left( 1 + \sum_{k \in N(j) \setminus \{i\}} m_{j\rightarrow k}(a - 1) \right).
\] (4.1)

In this case, we take the transmission probability from node \( i \) to node \( j \), \( \phi_{i\rightarrow j} \), to be inversely proportional to the degree \( k_j \) of node \( j \). We use a common global factor \( \phi \) for the entire network and take \( \phi_{i\rightarrow j} = \frac{\phi}{k_j} \).

We consider all of the same centrality measures that have been seen thus far in the thesis here. However, we also include another measure. We look at a centrality measure that like the BP centrality takes this inverse degree-dependence into account. We call this centrality measure the Inverse Degree-Dependent (IDD) centrality measure. The centrality \( x_i \) of a node \( i \) according to this measure is given by

\[
x_i = \sum_{j=i}^{k} \frac{1}{k_j},
\] (4.2)

where node \( i \) has \( k \) neighbours, each with degree \( k_j \). We decide to include this measure as it, like the BP centrality, ranks the nodes according to the inverse of their degree.

Before considering the performance of the BP centrality on undirected social networks, we look at its behaviour on an undirected synthetic network. We examine the BP centrality on the Synthetic Network 1 from Chapter 1. We saw previously that the measure performs very well when the dynamics on the network follow the simple SIR model with a constant transmission probability for each edge. We decide to test the BP centrality on the same network with the modified dynamics, where the constant probability of transmission from before is now the common global factor \( \phi \), with \( \phi_{i\rightarrow j} = \frac{\phi}{k_j} \). We believe this provides a smooth transition into the work that follows as we can immediately compare the performance of the measure on the original dynamics with its performance on the modified dynamics.
We know from Chapter 1 that the BP centrality performs very well on the Synthetic Network 1 in the sub-critical regime ($\xi < 1$), when a constant probability of transmission of $\phi_{i \rightarrow j} = 0.1$ is used across the network (Figure 3.9 and Figure 3.10). We can see clearly from Figure 4.1 above that the BP centrality also performs very well on the Synthetic Network 1 when the degree-dependent dynamics are considered with the non-uniform transmission probability of $\phi_{i \rightarrow j} = \frac{0.1}{k_j}$. The betweenness and IDD centralities are the next best predictors of highest influence.

We now look to see how the BP centrality behaves on some real-world social networks with non-uniform transmission probabilities. Using two networks we have already used, we run the modified SIR dynamics and test how well the BP centrality predicts the top influential spreaders in these real-world social networks.

We look at the well-known Karate Network which is previously seen in Section 3.5.2. The BP centrality performs well when we consider the original ICM with a constant transmission probability through the network in the sub-critical regime (Figure 3.18 and Figure 3.19). We now examine the BP centrality on the Karate Network again but alter the dynamics so that the transmission probability is inversely proportional to the degree of the node. To ensure the system is in the sub-critical regime with $\xi < 1$, we took the constant global factor of the network to be $\phi = 0.1$.
We can see clearly from Figure 4.2 above that the BP centrality also performs well when we consider the degree-dependent dynamics. This is a real-world social network with dynamics that are characteristic of real-world spreading phenomena. The results of the BP centrality in this situation leads us to suspect that this could be a promising measure. Also note that the IDD and degree centralities are the next best predictors.

To finish this section, we look at the Facebook Network previously examined in section 3.5.2. This is a real-world online undirected social network and to see how the BP centrality behaves when we alter the dynamics is of interest. The BP centrality performs well with the ICM where we consider constant transmission probability through the network (Figure 3.27 and Figure 3.28). We now consider the performance of the BP centrality on the same network but consider a degree-dependent transmission probability. We choose a common global factor of $\phi = 0.5$ to get a degree-dependent transmission probability of $\phi_{i\rightarrow j} = \frac{0.5}{k_j}$. 

Figure 4.4: Plots of the imprecision function and Jaccard distance for the Karate Network. Here we take $\phi_{i\rightarrow j} = \frac{0.1}{k_j}$, which ensures the system is in the sub-critical regime with $\xi < 1$. 

102
It is obvious from Figure 4.3 above that the BP centrality performs very well when we consider the modified ICM with transmission probabilities inversely proportional to the degree of the node. The next best performing centralities are betweenness and IDD. The fact that the BP centrality behaves well on this real-world undirected social network with dynamics that have a property of real-world social spreading processes is an encouraging result.

4.4 Results: Directed Networks with Degree-Dependent Dynamics

We now extend our work and look at directed networks. Recalling from Chapter 1, a directed network is a network with directed edges. If, for example, there is a directed edge from node $i$ to node $j$ but the link is not reciprocated, then there can only be a cascade in the direction from node $i$ to node $j$.

Social networks can be directed. Many online social media platforms such as Twitter and Instagram for example, are directed networks. On these apps, we can follow another person to see their content, but that person will not necessarily follow back all their followers. Thus, this creates a directed network of in-edges and out-edges from each node.

In this section, we examine the performance of the BP centrality on real-world directed networks with non-uniform probabilities of transmission. As previously discussed, the findings of [67] suggest a degree-dependent transmission probability on social networks, where the transmission probability $\phi_{i \to j}$, from node $i$ to node $j$ is inversely proportional to the in-degree $k_{(in)}j$ of node $j$. 

Figure 4.5: Plots of the imprecision function and Jaccard distance for the Facebook Network.
Here we take $\phi_{i \to j} = 0.5 \frac{k_j}{k_j}$, which ensures the system is in the sub-critical regime with $\xi < 1$. 

103
We use the ICM to simulate the dynamics, but unlike before we now take into account the direction of edges so that propagation from a node $i$ only occurs through the out-edges (as the direction of the edges is defined in a way that the arrows originate from the followee and point to the follower) of node $i$. Similarly to before, we also slightly modify the dynamics to account for these non-uniform probabilities of transmission. We use a common global factor $\phi$ for the entire network and take the independent edge transmission probabilities to be given by $\phi_{i \rightarrow j} = \frac{\phi_{(in)}}{k_j^{(in)}}$.

We evaluate the BP centrality using the same equation as before but take into account the direction of the edges. In this case we take $\phi_{i \rightarrow j}$ to be inversely proportional to the indegree of node $j$. The equation used to calculate the BP centrality in this case is given by

\[
m_{i \rightarrow j}(a) = \frac{\phi}{k_j^{(in)}} \left( 1 + \sum_{k \in N(j) \setminus \{i\}} m_{j \rightarrow k}(a - 1) \right)
\]  

(4.3)

As can be seen in equation (4.3) above, we use a transmission probability of $\phi_{i \rightarrow j} = \frac{\phi}{k_j^{(in)}}$, when calculating the BP centrality on directed networks.

We examine the performance of the BP centrality on three real-world directed social networks. Since we are dealing with directed networks, we consider centrality measures that are correspondingly appropriate for directed networks. In order to make a fair comparison between the BP centrality and the other measures, we believe it is necessary that these measures also account for the directionality of the network.

We look at the in-degree of the nodes as this could be an indicator of influence. In the case of social networks we believe that the in-degree could be an appropriate centrality measure. An individual’s influence on social media is not quantified by how many people they follow but the amount of people following them. On the other hand, we also consider the out-degree of the nodes here. We are attempting to identify the super-spreaders in the network. It seems reasonable to assume that nodes with a higher out-degree may better facilitate the spreading than nodes with lower out-degree. The closeness, betweenness and NB centralities here now take into account the direction of edges in the networks.

Again as in the undirected case, we consider a centrality measure that like the BP centrality involves an inverse degree-dependence. Like before, we call this centrality measure the IDD centrality measure. The centrality $x_i$ of a node $i$ according to this measure is given by

\[
x_i = \sum_{j=i}^{k} \frac{1}{k_j^{(out)}}
\]  

(4.4)

where node $i$ follows $k$ nodes, each with out-degree $k_j^{(out)}$. Note the slight difference between our definition for the IDD centrality in the undirected and directed case. This measure ranks the nodes according to the inverse of their out-degree. We include it here as it gives us another measure to compare with our BP centrality while also having an
Firstly, we examine the behaviour of the BP centrality compared to the other measures on the Physicians Network. This is a relatively small real-world directed social network, where directed links correspond to the respect of one member in the network to another member in the network.

Figure 4.6: Plots of the imprecision function and Jaccard distance for the Physician Network. Here we take $\phi_{i \rightarrow j} = 0.1 \frac{1}{k_{j}^{\text{in}}}$, which ensures the system is in the sub-critical regime with $\xi < 1$.

We take the common global factor of the network to be $\phi = 0.1$, which in turn gives a transmission probability of $\phi_{i \rightarrow j} = 0.1 \frac{1}{k_{j}^{\text{in}}}$. This choice of $\phi$ ensures the system is in the sub-critical regime with $\xi < 1$. We can see in Figure 4.4, that the BP centrality performs the best compared to the other centrality measures according to the results given by the imprecision function and the Jaccard distance. The out-degree is the next best performing predictor.

We now examine the performance of the BP centrality on the Wikipedia Network. This is a directed online social network. The links from user $i$ to user $j$ signifies that user $i$ voted for user $j$ to be a Wikipedia administrator.
Figure 4.7: Plots of the imprecision function and Jaccard distance for the Wikipedia Network. Here we take \( \phi_{i \rightarrow j} = \frac{0.01}{k_j^{(in)}} \), which ensures the system is in the sub-critical regime with \( \xi < 1 \).

Again, we choose a common global factor of \( \phi = 0.01 \), to ensure the process is in the sub-critical regime. This gives an edge dependent transmission probability of \( \phi_{i \rightarrow j} = \frac{0.01}{k_j^{(in)}} \) for the network. It is obvious from Figure 4.5 above that the BP centrality is the most accurate predictor of the top influential individuals in this social network, returning extremely low values for the imprecision function and Jaccard distance compared to the other measures. The next best centrality measures are the out-degree and NB.

Finally, we test the measure on a real-world Twitter Network. The analytical results of [67] are based on the social spreading phenomena of memes on Twitter. It is shown that the probability that a given meme is chosen for retweet is inversely proportional to the in-degree \( j \) of the node. It is of great interest for us to look at the performance of the BP centrality on a Twitter Network with non-uniform transmission probabilities as this relation resulted originally from work based on Twitter.
We used a common global factor $\phi = 0.15$. This ensures the spreading process is in the sub-critical regime. The overall edge dependent probability of transmission is given by $\phi_{i \to j} = \frac{0.15}{k_j}$. We can see in Figure 4.6 above, that the BP centrality not only performs very well but also performs significantly better than the other centrality measures. The BP centrality returns very low values for both the imprecision function and the Jaccard distance.

The BP centrality appears to be accurately predicating the ‘super-spreaders’ in directed social networks with dynamics that are characteristic of real-world social spreading phenomena. A spreading model with a transmission probability inversely proportional to the in-degree of the recipient node has been shown analytically and empirically [68]. The BP centrality performs well when these realistic (Twitter based) dynamics are considered on Twitter networks. For this reason, we are comfortable to propose the BP centrality as an accurate measure for the identification of influential nodes on social networks.
Chapter 5

Conclusion

Networks can be employed in a variety of ways. They are the backbone to the extensive comprehension of many real-world systems. Despite the diversity in the fields that exploit networks, the same questions and problems can arise across the board. The identification of influential nodes in a network is a constant challenge that we encounter when studying complex systems. It is a global problem that researchers face in a variety of fields. Be it sociologists analysing online social interactions, biologists investigating ecological network systems or even more recently in the new research domain science of science, physicists studying the roles of researchers in citation networks. The considerable demand for the answers surrounding influential node identification has motivated the work of this thesis.

To begin this thesis in Chapter 1, we introduce some basic methods and tools used in the field of Network Science. We include material necessary so to enable the uninformed reader to have a complete understanding and accessible comprehension of the new work and results provided later. In Section 1.2, we introduce some commonly used network models in the field, which we went on to use later. We discuss spreading models on network with an emphasis on the SIR disease spread model (Section 1.3) and discuss our process for the implementation of a variety of versions of this model. We include representations of the model that were based on discrete-time, continuous-time and percolation-based approaches.

In Chapter 2, we introduce a variety of currently existing centrality measures. We include degree, eigenvector, Katz, PageRank, closeness, betweenness, k-shell decomposition, and the more recent NB centrality (Section 2.2). We consider each of the measures individually, commenting on their advantages, where they produce accurate results and can well identify influential spreaders. We also mention their individual disadvantages and crucially note that the accuracy of a measure is completely dependent on ones definition of importance. Due to its relatively recent appearance in the literature, we dedicate Section 2.3 to the NB centrality measure. We comment on the literature available where it has been shown that the NB matrix has benefits over the adjacency matrix of a network. We also examine how it is inherently an extension of the well-known eigenvector centrality neglecting any localization effects. To finish this chapter, in Section 2.4 we give a brief review of the literature surrounding the area of ‘super-spreader’ identification that we believe to be prominent in the area.

We then introduce our newly proposed centrality measure based on a branching process approach, the BP centrality (Chapter 3). Again, for the unfamiliar reader we begin this chapter (Section 3.1) with a brief prelude on the theory of branching processes, including a discussion on probability generating functions and Galton-Watson processes. In Section 3.2, we present a mathematical model based on a branching processes approach for the predication of influential nodes in a network, which we define as the BP centrality measure. We introduce the mathematical model and explain its innate assumption that
the network under examination is tree-like due to it being based on a branching process model. We expand on this and go through the implementation of the BP centrality measure. We acknowledge the possibility in theory of calculating the BP centrality by solving the linear system but, use an iterative method in our work for computational efficiency. In Section 3.3, we refer to Section 1.3 and explain the implementation of spreading dynamics we use for our work. We introduce the imprecision function and Jaccard distance which are used to quantify the accuracy of the centrality measures. We dedicate Section 3.4.3, to the network data and sources used for the entire work of this thesis. We also include a section on the computational efficiency of the BP centrality measure. We conclude the section with some mathematics on the performance of the NB and BP centralities at criticality. We refer to the literature and extend on the work found there.

We present the performance of the BP centrality and a variety of other centrality measures on a multitude of networks (Section 3.4). We begin with some synthetic networks based on models from Section 1.2 and then, present results on some real-world networks. Having examined the BP centrality measures performance on networks with arbitrary parameters, we go on to investigate its behaviour in the sub-critical, critical and super-critical regime on synthetic and real-world networks. The BP centrality assumes the network is tree-like with children nodes independent from each other. However, many networks have short loops such as triangles. In the sub-critical regime, we are able to see from our results that the existence of these clusters in the network do not affect the accuracy of the BP centrality. However, in the super-critical regime, the BP centralities performance tends to be very poor. In this scenario, the assumptions of the branching process model are violated. With extensive spreading throughout the network that occurs in the super-critical regime, the assumption of the tree-like structure of the network no longer holds. To finish Chapter 3, we present the performance of the BP centrality on some real-world social weighted networks where the weights correspond to time spent between individuals. For a fair comparison here, we examine the performance of the BP centrality with some weighted centrality measures.

In the final chapter of this thesis, we consider the performance of the BP centrality when we deal with degree-dependent dynamics (Chapter 4). We explain the theoretical and empirical work which finds the probability of transmission between nodes on social networks to be inversely proportional to the degree (in-degree) of the recipient node. Using undirected and directed real-world social networks, we examine the performance of the BP centrality with degree-dependent transmission probabilities on social networks with degree-dependent spreading dynamics. We found that in sub-critical regime, the BP centrality performs very well on social networks with dynamics that follow this degree-dependence.

In this thesis, we present a new centrality measure, the BP centrality. We test its accuracy on a variety of networks. From the extensive examination of the theory and our results, we are satisfied to propose the BP centrality as an accurate measure for the identification of influential nodes in tree-like networks. Further to this, the measure performs very well on networks that are not necessarily tree-like in the sub-critical regime. On social networks, where the dynamics are generally sub-critical and the transmission probabilities are inversely degree-dependent, the BP measure also produces very accurate results. At
criticality, we show theoretically and by example that the NB and BP centrality coincide. As previously discussed in Section 3.4.5, the NB centrality has been proven to be the optimal measure for tree-like networks at criticality. The BP measure provides the same accuracy here and so could be another possible tool for the identification of influential nodes in these scenarios.

Despite the impressive accuracy of the BP centrality in certain situations, it is not a globally superior measure. The BP centrality is based on a branching process mathematical model and assumes children nodes in the network are independent. For this reason, the measure performs poorly in the super-critical regime where the spreading through existing loops in the network is imminent. For this reason, we do not suggest our measure to be used in these situations.

In this thesis, we have presented a new centrality measure and shown through our extensive range of results its advantages and disadvantages. However, there is still a lot of work to be undertaken in this space. In our work, we consider the influence of node in the limit of large time when the system settles down and the spreading dynamics stop. We rank nodes according to their spreading power in the steady state. A possible direction for future work could be the time-dependent BP centrality measure. An important question encountered could be: What node should we infect in order to have the largest outbreak within the first 10 days of infection? This node may not be the most influential in the steady state. This opens the question of time-dependent centralities and the changing of node rankings with time. This could be important in situations of time constraint like for example the spread of information in a crisis or the isolation of an individual to inhibit the spreading of a life-threatening disease within a community.

Another avenue to consider is the performance of the BP centrality in the area of collective influence. In some situations, we do not necessarily need to know the of most influential individual seeds in a spreading process. We may want to know the groups of most influential seeds that we should infect simultaneously to maximize the outbreak. This could be of interest to companies that are looking to advertise their products through social media influencers, for example. The companies want to strategically choose a set sized group of influencers to promote their product, with the intention that the group will reach the largest audience and also be influential enough to sell the product to their followers [76][77].

In summary, we provide a new centrality measure called the BP centrality for the identification of influential nodes in a network. We discuss and show that this measure is most accurate on tree-like networks, networks with dynamics taking place in the sub-critical regime and on social networks with degree-dependent dynamics. We reinforce our suggestions regarding its accuracy and failures by presenting an abundance of examples. We welcome further work and the development of the BP centrality in future research.
References


