Imperial College London

BEng Individual Project Report

Department of Computing

Imperial College of Science, Technology and Medicine

Modelling the Effects of Social Products and Other Factors on Information Flow within Social Networks

Author:
Thomas Claydon

Supervisor:
Alex Carver

Second Marker:
Susan Eisenbach

June 14, 2020

Abstract

Increasingly in recent years individuals have turned to online social media and social network services for the purpose of news consumption. This trend has brought increased focus and attention to the fields of social network analysis and modelling. Many existing models for information diffusion have focussed on the interactions between individuals and news sources, whilst placing little emphasis on the fundamentally social nature of modern online platforms.

The main goal of this project is to develop a new model for information cascade that places the *social* interactions between individuals at the centre of its dynamics. In order to do this we thoroughly explore existing models and use their findings to influence our design.

We build our information cascade model using an iterative process, at each step evaluating our model both qualitatively and quantitatively using the metric of "mean edge homogeneity" from the literature. We find that our model behaves similarly to traditional models, highlighting the important role of social interactions in information diffusion and validating our hypothesis. We then carry out an investigation into the role of other factors that contribute to information diffusion including network topology and choice of seed node. We find that node opinion has the largest impact on cascade dynamics out of those factors tested. As part of this experimentation we also found that by distributing node opinion bimodally we were able to observe results and behaviours even closer to real life studies.

Acknowledgements

I would like to thank my supervisor Alex Carver for his constant support and guidance throughout the completion of this project and particularly throughout the challenging and unprecedented events of the last few months. Additionally I would like to thank my second marker Susan Eisenbach for her valuable feedback during the early stages of the project.

More generally I would like to thank all of the people, both staff and students, that I have met throughout my time at Imperial.

Finally I have to say a massive thank you to all of my incredible friends and family for their unwavering support throughout the years. In particular my mother, father and grandparents have provided continuous love and guidance, and I would not be where I am without them.

Contents

Chapter 1

Introduction

1.1 Information Spread Online

The internet has brought about a seismic change in the way individuals communicate information and ideas. In recent years, media consumption and behaviour has evolved and the advent of online social media services has redefined how individuals receive and share news. This new approach to information exchange through online social networks allows individuals access to a wider array of news sources in an on-demand fashion, and has introduced an interactive element to a previously static process: this has drastically altered the worldwide information ecosystem.

Using online social networks individuals are able to publish their own thoughts and opinions directly with others with incredible speed and exceptionally low cost or effort required, creating seamless interactions between content producers and consumers which has rapidly increased the speed at which all forms of information can spread online. This behaviour is aided by a lack of standards and regulations surrounding social media, and the fact that many platforms "succeed" by actively discouraging users from substantiating claims [1].

This disintermediation of the way in which individuals consume news makes it harder for claims to be verified, and as a result unsubstantiated rumours can spread quickly. At the same time recent advancements in AI technologies, particularly machine learning, have allowed for the faster creation of misinformation in ever larger quantities and at higher qualities, making it even harder to detect its deceptive nature. As a result, the World Economic Forum has categorised "massive digital misinformation" as a central technological issue that could "wreck havoc in the real world" [2].

The potential damage to society that can be caused by the spread of unsubstantiated rumours (more colloquially referred to as "fake news") is particularly highlighted during times of crisis, such as during the ongoing COVID-19 pandemic of 2020: in fact the pervasiveness of misinformation online throughout the pandemic has led some researchers to study the role that the ongoing "infodemic" has had on the effectiveness of the response to the public health crisis [3]. In situations such as this, it is not hyperbole to suggest that misinformation can result in physical harm to society and even cause death to individuals [1].

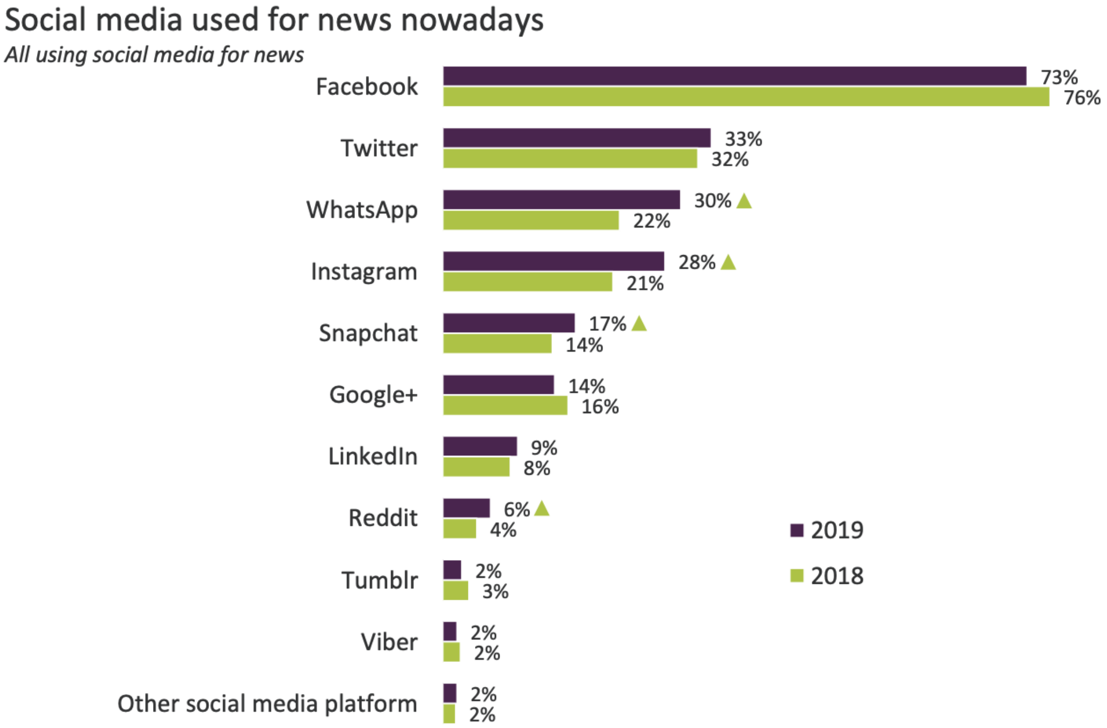


Figure 1.1: How different online platforms are used for news consumption within the UK amongst the 49% of adults that claim to use social media for news, as shown in [4, p. 41].

As the world becomes ever more interconnected thanks to technological progress, it becomes increasingly important to understand the causes of certain information flow phenomena. Having discussed the possible dangers and implications of the internet's effects on information spread, we can see a clear similarity between how ideas can spread online and medical epidemiology: this field has come to the forefront of most individual's attention during the ongoing 2020 COVID-19 pandemic. As has been widely reported, a key aspect of the international response to fighting a physical virus is being able to model how it is currently spreading and will spread in the future through different populations under different situations. Similar tools and techniques can be equally important in the fight against (mis)information spread online, helping to inform decisions to combat its effects going forward.

Social media allows for information of all kinds to spread at an incredible pace, compared to traditional forms of media. Direct access to such a broad spectrum of information naturally leads to individuals selecting information that is compatible with their interests and beliefs. This tendency for an individual's online profile to become ever more personalised over time thanks to the site's "algorithm" can distort the individual's world view: this becomes all the more important as social media takes over from traditional mediums as the main source of news. As of 2019, 49% of adults in the UK consume the majority of their news via social media [4]. Figures 1.1 and 1.2 highlight how the majority of respondents primarily consume their news from social media posts rather than directly from verifiable news sources, and additionally show that 3 in 4 of those using social media use Facebook as their primary choice of social media for news consumption.

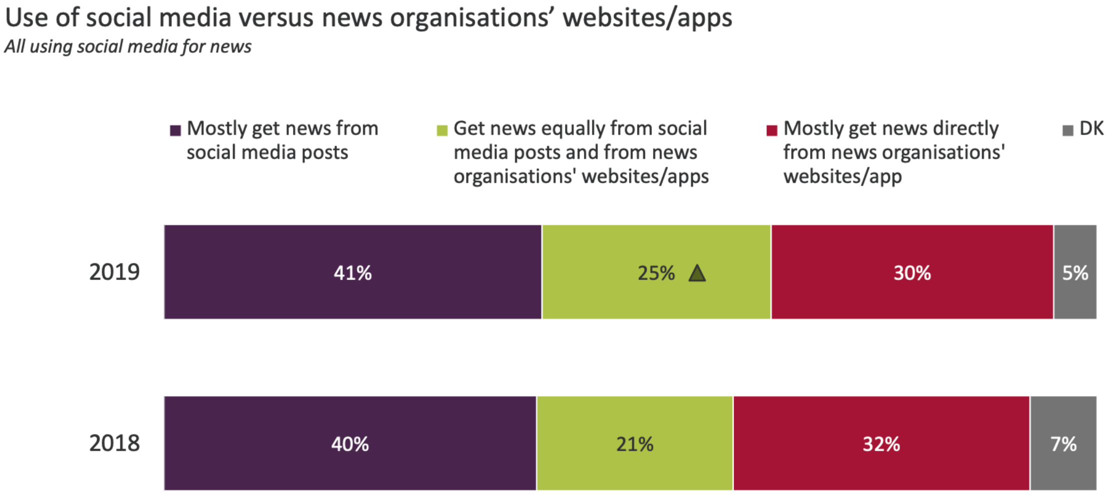


Figure 1.2: A breakdown of how the half of UK adults who use social media for news consumption use other sources as shown in [4, p. 43].

In addition to the highly personalised nature of social media feeds, the interactive nature of online platforms combined with the ability to have quick and easy access to like-minded individuals allows ideas to spread socially through "spheres of influence" far easier than through traditional in person social networks. In fact such a personalised experience online has been shown to foster the formation of *echo chambers* [5] (the tendency for individuals to communicate in clusters with others of similar views), whilst Vicario et al. have studied the role of confirmation bias in this phenomena [6]: this will be explored in section 3.1. These findings feed a central hypothesis of this work: nodes in a social network have a preference over who they communicate and share information with. It is these social aspects to the sharing of information online that motivate our work, and differentiate it from the literature.

1.2 Objectives

In this project, we aim to create a new model for information flow within social networks. The core of our model will be based on the premise that individuals have pre-existing beliefs and preferences over who they share information with. This is largely influenced by existing independent cascade and threshold models for diffusion which are discussed in section 2.4, and recent research into the real world factors affecting information cascade dynamics discussed in section 3.1.

We aim to evaluate our model by verifying its behaviour both qualitatively and quantitatively against that observed in real social network studies in previous work: specifically that is the tendency for individuals to communicate in clusters of similar opinions.

Then using our model, we aim to carry out an investigation into the way in which different factors affect information diffusion. Such other factors include:

- social network size;
- network topology;

- “opinion” distribution;
- seed node choice or distribution.

This investigation will require an informal “pipeline” for the design, execution and analysis of experiments that include the simulations of many information cascades across many social network structures with differing parameters. This will allow us to thoroughly evaluate our model across a wide range of different scenarios and better relate our work back to the wider literature.

As part of our model evaluation and investigation we will require methods for generating and analysing social networks with varying properties. To best test our model performance on real social networks, we will want to include at least one real world social network structure in our investigation. We aim to include some kind of interactive visualisation of cascades as part of the analysis to best show how the network evolves throughout the simulation, however this interactive “application” is not the core focus of our project and as such is regarded as a secondary goal.

1.3 Challenges

Throughout the completion of the project there are a number of key challenges and milestones that we anticipate.

Understanding of Diffusion Models To inform our model design and ensure it provides a new viewpoint to questions and topics covered in the literature, we must perform thorough background research and analysis of existing diffusion models and techniques used in the field of social network analysis. This knowledge will help us to shape our model and subsequent investigation.

Real-World Network Example Availability As discussed in section 1.2, we aim to include some examples of real online social network topologies in our experimentation to help us evaluate its usefulness and effectiveness in modelling information flow on real social networks. Therefore we will require some pre-existing dataset that encodes social network examples from real online platforms, or we may need to devise our own method for obtaining this data.

Quantitatively Measuring and Evaluating Model Success The primary method of evaluating the “performance” of our model is through the observation of behaviour similar to that seen in real social networks described in the literature. Qualitatively this is reasonably simple: through visualisations we can observe the network topology and the behaviour of the information cascade to see whether information “spreads” within clusters. However being able to introduce a quantitative metric to more precisely measure this behaviour would be beneficial as it will provide a more straightforward way of comparing and analysing experiment results.

1.4 Contributions

We explicitly summarise our contributions that will be presented in detail throughout the remainder of this report.

- A thorough and self-contained summary of existing graph and network theory and state-of-the-art models for diffusion within social networks, presented in chapter 2.
- An explanation of research and studies into the real world factors and behaviours that affect information cascade dynamics, presented in chapter 3.
- A new model for information cascade focussing on *social* effects, that takes individual's pre-existing beliefs and resulting preferences into account, presented in chapter 4.
- An analysis of our models behaviour in relation to the real world observations from the literature, as well as the results and analysis of an investigation into the role of other factors in information cascades online, presented in chapter 6. As part of this experimentation work, we also provide a new and alternative method for generating directed scale-free graphs.

Chapter 2

Background

2.1 Social Networks

The term *social network* is used to describe a set of entities and the interactions and connections between them. In modern society the term is most often used colloquially to describe online social network applications[1] such as Facebook, Instagram or Twitter, however the term is equally applicable to traditional in-person networks such as a group of friends or a family.

In recent years as the internet has become more pervasive and online social networks have become more prevalent, research in the area of social networks has grown to become highly interdisciplinary. A key focus of research has been on analysing *diffusion*, which describes how an entity (idea, piece of information, event, physical product, etc.) spreads across a network. In this work, we focus on the diffusion of information and social products, and the interplay between them. These notions are discussed in section 2.3.

2.2 Graph and Network Theory

For analytical purposes social networks are most commonly represented as mathematical graphs. At the most simplistic level a social network can be represented by a graph $G = (V, E)$ where the set of vertices V represents the entities[2] of the social network (people, social groups, etc.) and the set of edges E contains pairs of vertices (i, j) where $i, j \in V$, representing the social connections[3] between entities. We often consider the graph G to be *simple* i.e. to contain no self loops of the form $(i, i) \in E$ and no parallel edges such that there exists at most one edge $(i, j) \in E$ between nodes i and j. We denote the cardinality of sets by $|\cdot|$: for example, we often consider social networks with n nodes, where $n = |V|$.

Such simple graphical representations for social networks are highly versatile, as they are highly extensible when creating more complex models as we shall discuss in section 2.4.

[1] We use the term "social media" to collectively refer to examples of such online social network applications throughout this report.

[2] Throughout this work and the literature, many terms are used to refer to the set of entities V in a network or graph: these include common terms such as "node", "vertex" and "agent" as well as more specific terms for the context of human social networks such as "user" and "individual".

[3] Most commonly such social connections refer to friendships, either in the more traditional sense or the more modern interpretation from online social networks: in this situation the term "follower" may also be used, which normally refers to a one-way relationship leading to the necessity for directed graphs.

2.2.1 Graph Terminology

To allow us to accurately describe and classify different social networks throughout our investigation, we now introduce some fundamental terminology in the field of graph and network theory.

Directed and Undirected Graphs

For any diffusion in a social network to be modelled, we require a notion of *relationships* between nodes in the graph. Such relationships can be regarded as either directed or undirected, depending on the real-world context of the relationship being modelled.

In an undirected graph, an edge $(i, j) \in E$ between nodes $i, j \in V$ denotes the existence of some bidirectional relationship or interaction between nodes i and j: we refer to node j as a *neighbour* of node i. The set of all neighbours of i is denoted by $N(i)$, and the cardinality of this set $|N(i)|$ is defined as the *degree* of node i, denoted by d_i. Such bidirectional relationships are typical in in-person social networks and some online social networks such as Facebook: when you become "friends", the connection works both ways allowing for communication or interaction in both directions.

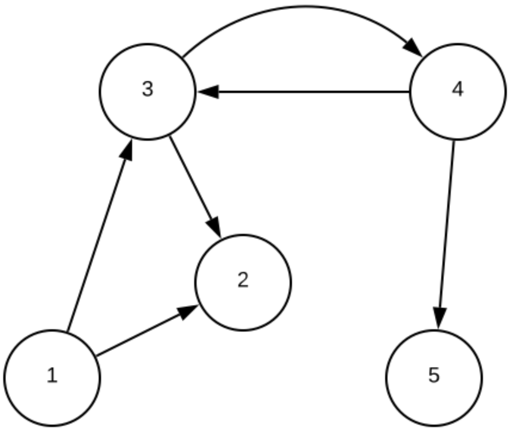
When working with directed graphs, we can extend our notions to take account of the directed nature of the edges. An edge $(i, j) \in E$ between nodes $i, j \in V$ in a directed graph $G = (V, E)$ represents an *outgoing* relationship from node i to node j and so node j is defined as an *outgoing* neighbour of node i. We denote the set of outgoing neighbours for node i by $N^{out}(i)$, and the outgoing degree of i by d_i^{out}. Similarly, for a given node $i \in V$ we define the set of incoming neighbours as all nodes $j \in V$ such that there exists an edge $(j, i) \in E$ and denote this set by $N^{in}(i)$: again the cardinality of this set is then defined as the incoming degree of i, denoted by d_i^{in}. Note that we can still apply the general definitions of neighbours and degree to directed graphs if we consider them as undirected, however this is uncommon. Examples of unidirectional relationships include the "following" system on some online social networks such as Twitter and Instagram: you can "follow" another individual without them having to "follow" you back. Figure 2.1 graphically shows examples of these definitions.

Degree Distribution

As we have just defined, the degree d_i of a node i in a graph denotes the number of the neighbours node i has: mathematically that is the cardinality of the set of all neighbours of i. In isolation, considering the degree of singular nodes in a graph is rarely of much analytical interest. Instead it is much more useful to consider the *distribution* of node degrees in the graph: that is the probability distribution of the degrees of every node across the whole graph or network. A histogram or probability density plot of the degree distribution can be helpful in identifying and classifying network structures, as we will discuss in section 2.5.

Connectedness: Weakly and Strongly Connected Components

An undirected graph is *connected* if there exists a path between every pair of nodes in the graph: if not connected as a whole, a graph is made up of a number of *connected components*. When working with directed graphs, we can extend this notion: a directed graph is *strongly* connected if there exists a path between every two nodes in the graph,



$$
\begin{aligned}
N^{in}(1) &= \emptyset \\
N^{out}(1) &= \{2,3\} \\
N(1) &= N^{in}(1) \cup N^{out}(1) \\
d_1^{in} &= 0 \\
d_1^{out} &= 2 \\
d_1 &= d_1^{in} + d_1^{out}
\end{aligned}
$$

Figure 2.1: A simple example of a directed graph and examples of the application of the terminology and notation introduced in 2.2.1.

whilst a directed graph is *weakly* connected if it has the property of every node being reachable from every other node if we ignore the directional element of the edges. For example, the simple directed graph shown in figure 2.1 consists of a single *weakly* connected component: it is not strongly connected as there is no path between every pair of nodes when the edges direction is enforced.

The notion of a graph's *connectedness* is useful when considering the interpretation of some graph metrics that we will introduce in section 2.2.2, as well as in the discussion of differing network topologies and real-world datasets as we will do in sections 2.5 and 3.2 respectively.

2.2.2 Graph Metrics

When discussing different social networks throughout our investigation, we will require a number of metrics to allow us to compare and categorise both quantitatively and qualitatively different networks to each other. In this section we present a number of graph metrics that will be useful in the development of our model and subsequent discussion and analysis of our cascades.

Average Clustering Coefficient

Much of the motivation for and focus of this project revolves around the observation that in real-world social networks people tend to cluster together and form *cliques*[4]: the average clustering coefficient of a graph helps us to quantitatively define this. In the literature there are many different ways in which the clustering coefficient is defined. Throughout this work we define the average clustering coefficient as follows, as first presented by Watts and Strogatz in [7]:

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{2.1}$$

[4] A clique C in a graph $G = (V, E)$ is a subset of nodes $C \subseteq V$ such that every pair of nodes in C are directly adjacent to each other i.e. for every pair of nodes $i, j \in C$ there exists an edge $(i, j) \in E$.

where each C_i is the *local* clustering coefficient of node i. The local clustering coefficient of a node i represents how close its set of neighbours is to being a clique: this is the fraction of actual links between its neighbours, divided by the number of possible links. More precisely, Watts and Strogatz define the *neighbourhood* N_i of a node i to be the nodes that it is connected to: $N_i = \{v_j : e_{ij} \in E \vee e_{ji} \in E\}$. They then define k_i as the number of nodes in the neighbourhood i.e. $k_i = |N_i|$. For a node i in an undirected graph, there can be at most $\frac{k_i(k_i-1)}{2}$ edges between the k_i nodes in the neighbourhood if it is fully connected. Therefore for an undirected graph $G = (V, E)$ each C_i for $i \in V$ is defined as:

$$C_i = \frac{2\,|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)} \tag{2.2}$$

and similarly for a directed graph, due to the possibility of each pair of nodes in the neighbourhood being connected in both directions, we simply ignore the factor of two in the numerator. The clustering coefficient of a network is one of the key metrics used to classify networks into different topologies with different properties as we shall discuss in section 2.5.

Transitivity

Transitivity is a closely related metric to average clustering coefficient. It differs in that it considers the entire network at the macro level. The global clustering coefficient of a graph G is defined in [8, p. 243] as:

$$C = \frac{\text{number of closed triplets}}{\text{number of all triplets}} = \frac{3 \times \text{number of triangles}}{\text{number of all triplets}} \tag{2.3}$$

where a triplet is three connected nodes (open if connected by two edges or closed if connected by three edges). Both transitivity and average clustering coefficient provide a quantitative sense of the clustering within the network, however the different scales of focus provide different information on the structure of a network.

Average Shortest Path Length

As the name suggests, the average shortest path length is the average of the shortest path between every two pair of nodes in the network. For a graph $G = (V, E)$, this is:

$$a = \sum_{i,j \in V, i \neq j} \frac{d(i,j)}{n(n-1)} \tag{2.4}$$

where $d(i, j)$ is the shortest path between nodes i and j (each edge contributes one to the distance) and $n = |V|$. We say that $d(i, j) = 0$ if node j cannot be reached from node i, such as if the network is not connected. For directed graphs, we can calculate the average short path length if the graph is weakly connected, by ignoring the directional element of the arcs. Along with the average clustering coefficient, the average shortest path length is the most important and distinguishing metric for the classification of networks.

Diameter

The final metric that we shall describe is the diameter of a graph. The diameter of a graph $G = (V, E)$ as defined in [9] is the longest shortest path between any pair of nodes:

$$d = \max_{v \in V} \max_{u \in V} d(v, u) \tag{2.5}$$

that is, for each node find the *shortest* distance between it and every other node and record the largest of these values, and then select the largest of these values across every node in the graph. This tells us the maximum number of arcs between any two nodes in the graph, and alongside the average shortest path length provides insight into how the graph can be traversed.

2.3 Information and Social Products

In section 2.1 we said that this work will focus on diffusion in social networks in the context of information and social products. These two concepts are highly related, and we now provide an overview and intuition for the differences between them and their importance for our work.

Information We consider information in the context of social network analysis to be a piece of news or knowledge that can be received, and then passed on. This definition includes both facts and unsubstantiated claims. From [10], *misinformation* is a type of information that is false or inaccurate: note that we make the distinction that misinformation does not have to be malicious in intent. Meanwhile *disinformation* is a subset of misinformation with the specific and deliberate intent to deceive [10]. At a high level information flow across a network can be analogous to the spread of a biological virus or contagion. It can be transmitted from node to node, and can result in a change of state at each node: the nature of the change of state is where the analogy starts to weaken. Whilst a virus simply transmits through an individual, causing a change of state from “healthy” to “unhealthy” as a result, information must be *adopted* by the individual to cause a state change. If this adoption occurs, it can possibly cause a change to the individual’s *social products*.

Social Products As introduced above social products describe a change in the state of a node in a network that occurs as a result of some social interaction. For example entities such as political opinion, religious beliefs and group membership are common examples of social products: nodes can have different political opinions or be members of different groups or societies, and these can change dynamically as a result of social interactions and experiences. The key property linking these intangible entities as social products is that they are transmitted socially and must be adopted by the individual.

Information and social products are clearly interrelated: information may be shared and potentially adopted by users in a social network depending on their pre-existing beliefs i.e. social products, and adoption of certain beliefs may also influence an individual’s overall belief system inducing a change in their social products. This circular nature of the relationship between the two related notions is what has led to the creation of different models, theories and mechanisms to study the ideas separately as it becomes complex to study both phenomena at the same time. In section 2.4.2 we introduce independent cascade models which are commonly used in the study of information flow and in section 2.4.3 we introduce a threshold model more appropriate for the modelling of social product diffusion.

From this it is clear that pre-existing social products in a social network will have a large influence on information flow. For example different people in a social network will have different beliefs and opinions. In section 3.1 we present work from the literature that analyses interactions within real social networks and find that individual’s are more likely to communicate with others that are like-minded. This is a central hypothesis of this project: pre-existing social products in a social network influence how information is diffused across the network.

2.4 Existing Diffusion Models

In this section we discuss the most common categories of models for diffusion in social networks, and present some existing models that have motivated and inspired our work.

Many early diffusion models came from the field of epidemiology, as researchers looked to model the spread of contagions through a community. Over time as the applications of social network analysis broadened, these early epidemiological models were adapted to better reflect the situation in question.

2.4.1 Susceptible-Infected-Recovered (SIR) Models

An early, seminal model for diffusion was the SIR model that was introduced in [11]. SIR came out of the field of epidemiology, and the influence of this field can be clearly seen in the design of the SIR model. At its simplest, SIR models consider each node in the network to be in one of three states:

- susceptible, meaning the node can be infected by a neighbour; or
- infected; or
- recovered, meaning the node is no longer infected and cannot be reinfected (either temporarily or permanently depending on the exact situation being modelled).

The diffusion starts from a set of one or more seed nodes, each of which can "infect" each of its neighbours *who are in a susceptible state* with a parameterised probability p. At each advancement of time, any node that was in an infected state at time $t-1$ transitions to a recovered state.

The limiting factor of the SIR model is the single infection probability p. The use of a single, global probability for all nodes and diffusion steps in the network is a strong generalisation that may not always provide an accurate representation of the situation being considered. This led to the development of more generalised models, such as independent cascade models, which we will discuss in section 2.4.2.

2.4.2 Independent Cascade (IC) Models

Independent cascade models such as that discussed in [12, p. 35–36] are generalisations of the SIR model discussed above in section 2.4.1. Like SIR, independent cascade models use discretised time steps to capture the diffusion process, however nodes can only be in one of two states: active or inactive. Diffusion is modelled through the tendency of inactive nodes to become activated probabilistically through interactions with neighbouring nodes.

As discussed in section 2.4.1, SIR models use a single probability of infection for the entire network, whilst IC models allow for the specification of infection probabilities for each edge in the network. An edge $(i, j) \in E$ has an associated probability $p_{i,j}$, representing the likelihood of node i activating node j at a given step of the diffusion process.

Such a generalisation of the infection probabilities allows for the model to be tailored to fit a wider range of real world scenarios. For example, the probability $p_{i,j}$ can be assigned using a number of real world factors, such as geographic location or previous infection rates. Once a node is activated, it activates each of its neighbours at the next time step, according to the probabilities associated with each edge. The important factor that makes the cascade *independent* is that each node has only one chance to activate each of its neighbours, at the next time step after it was activated itself.

Figure 2.2 shows a simple example of an independent cascade model. Initially at time $t = 0$ the seed set contains active nodes C and D, highlighted in yellow. At time $t = 1$, nodes C and D are able to activate their respective outgoing neighbours (A, G and H) and (B, E and F) with the probability specified adjacently to the edge: subsequently nodes A, E and H are infected and become activated, and the previously active nodes C and D remain active but are unable to further infect any neighbours. The cascade continues until $t = 3$ when there are no longer any nodes that can be activated, and as a result the diffusion stops.

This example assumes that a node can only transition from inactive to active, after which it will remain active for the remaining duration of the model: as a result the term "monotonic" is occasionally used for models under such assumptions. Such an assumption is useful when modelling the diffusion of information, as it is unintuitive to suggest that nodes can "forget" information once they have received it. However, when considering the diffusion of other entities such as *social products* as we will see in section 2.4.3, such an assumption is restrictive as social products can change over time.

2.4.3 Threshold Models

Linear Threshold Models

Standard linear threshold models have a similar structure to the independent cascade model discussed in section 2.4.2, however the diffusion dynamics differ. We now discuss an example of a linear threshold model as shown by Shakarian et al. in [12, p. 38]. Each node can again be either active or inactive, however additionally associated with each edge $(i, j) \in E$ is a non-negative weight $w_{i,j} \in [0, 1]$. It is also assumed that for any node the sum of the weights of incoming edges is less than or equal to one i.e. for each node $i \in V$ that $\sum_{j \in N^{in}(i)} w_{j,i} \leq 1$. A threshold function $\theta : V \to (0, 1]$ is then defined that assigns a value in the interval $(0, 1]$ to each node $i \in V$.

At a time t, each node that was active at time $t - 1$ has the opportunity to activate its inactive outgoing neighbours. Intuitively, a node will become active if the sum of the weights of its incoming edges from previously activated nodes is greater than or equal to its threshold. More formally an inactive node i will be activated at time t of the diffusion if:

$$\sum_{j \in N^{in}(i) \cap H_{t-1}} w_{j,i} \geq \theta(i) \tag{2.6}$$

where H_{t-1} denotes the set of all nodes that were active at time $t - 1$.

The interesting difference between independent cascade and threshold models is in the idea of collective social influence. Independent cascade models allow for a single node to

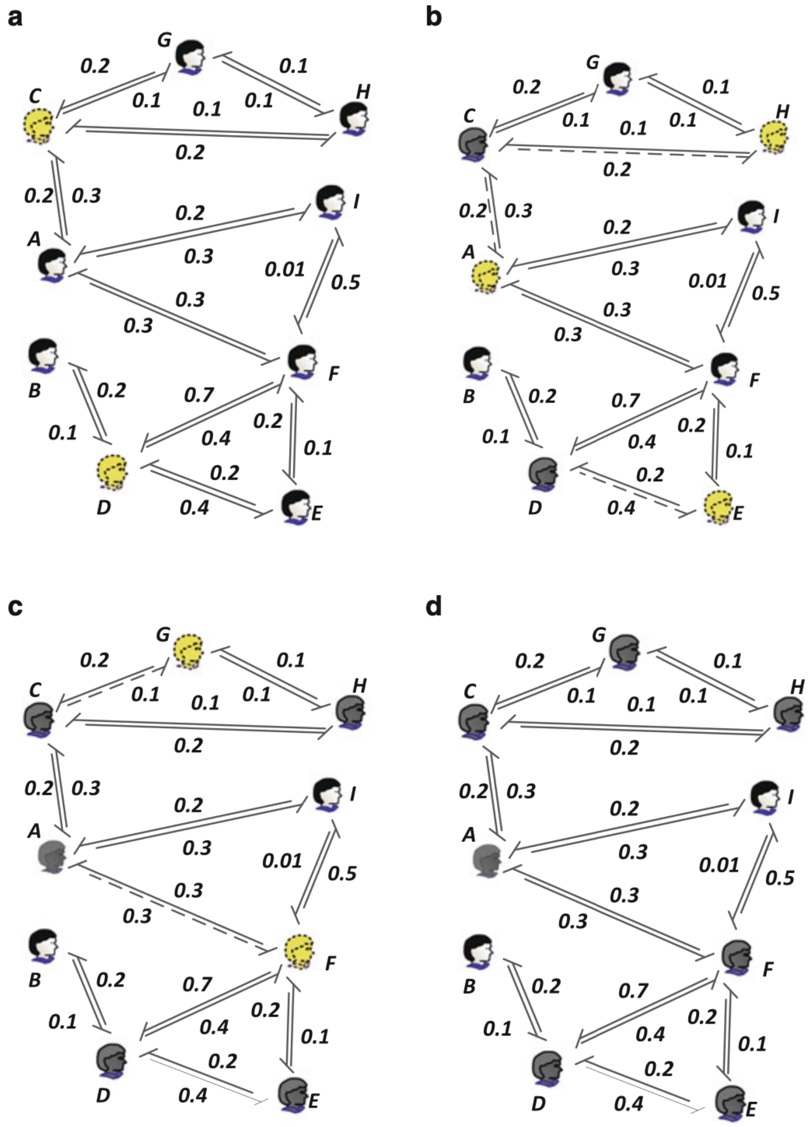


Figure 2.2: An example of an independent cascade model progressing from $t = 0$ to $t = 3$ as shown by Shakarian et al. in [12, p. 37]. Section 2.4.2 describes the figure in more detail.

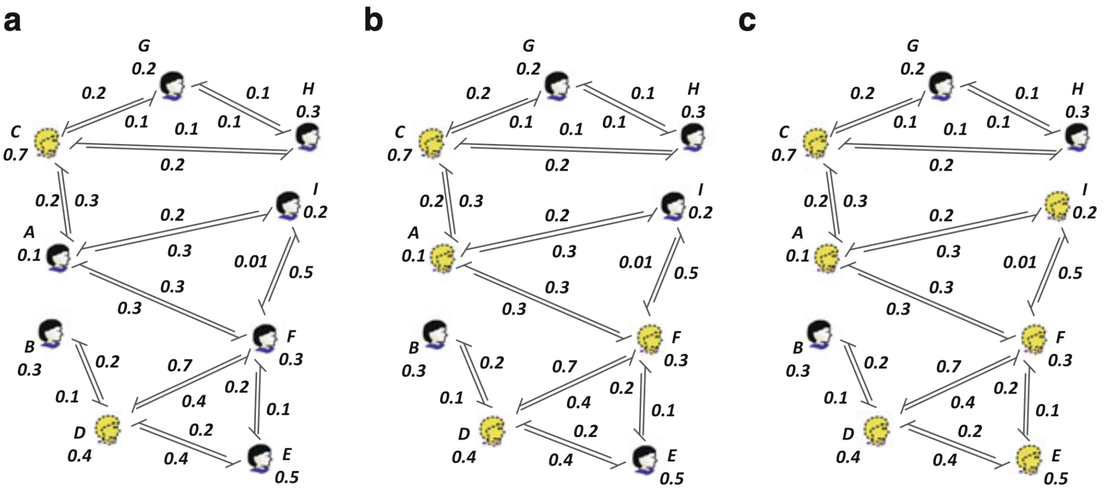


Figure 2.3: An example of a threshold model for diffusion across a social network, progressing from $t = 0$ to $t = 2$ as shown by Shakarian et al. in [12, p. 38]. Section 2.4.3 describes the figure in more detail.

activate or "infect" its neighbours (probabilistically), whilst threshold models more closely capture the often collective and social aspect of diffusion: nodes can only be activated when *collectively* its neighbours are able to combine their influence to exceed the nodes threshold. Such behaviour is representative of how ideas often spread online, as we shall discuss in section 3.1.

Figure 2.3 shows an example execution of the threshold model we have previously described. In this example from [12, p. 38], the threshold function θ assigns thresholds to nodes randomly and uniformly in the interval $[0, 1]$. This is done intentionally for this simple model as there is no pre-existing knowledge of the behaviour of nodes in the network. In more advanced models, such thresholds can be assigned to capture real-world behaviours and model the preferences and social products of nodes.

When $t = 0$, nodes C and D are active, and so at $t = 1$ they have the opportunity to activate their inactive neighbours. Node C is only able to activate node A as $\theta(A) = 0.1$ and $w_{C,A} = 0.2$ i.e. $0.2 \geq 0.1$ so the influence of node C exceeds the threshold of node A. Note that as C is the only active neighbour of A, it is only able to influence A on its own. Similarly, D is able to activate F as $0.4 \geq 0.3$.

At $t = 3$, there are now four active nodes in the network. Collectively they are able to activate node I as $0.5 + 0.3 \geq 0.2$, and node E as $0.2 + 0.4 \geq 0.5$. This example highlights how nodes can collectively influence common neighbours in threshold models.

2.5 Network Topologies and Generation

Our mechanism for generating cascades will be built on top of underlying graphs that represent real-world social networks of individuals. In this section we discuss three commonly discussed network topologies, and well-known models from the literature for generating such graphs which will become important during our model experimentation in chapter 6.

2.5.1 Erdős-Rényi Random Graphs

Random graphs are those in which the edges between pairs of nodes are distributed randomly. The Erdős-Rényi model was presented in [13] as a new model for the generation of random graphs. An Erdős-Rényi (ER) graph $G(n, p)$ is parameterised by the number of nodes n and the probability of edge inclusion p. Edges are added between pairs of nodes randomly and independently of each other with probability p, resulting in a graph with $\binom{n}{2}p$ edges on average.

The base assumptions of the Erdős-Rényi model (that edge inclusions are independent of each other and equally likely) are unrealistic in most real world settings: this leads to ER graphs having low clustering coefficients and larger average shortest path lengths than many real-world graphs. However, these properties and the stochastic nature of the graph generation means that ER graphs can often still be a useful benchmark to compare against when considering graph metrics to other topologies.

2.5.2 Small-World Networks

The small-world phenomena, or more colloquially the principle of "six degrees of separation", suggests that almost every pair of individuals in a human social network is connected by a short path [14]: intuitively, most individuals are not friends with each other, but your friends are likely to be friends with each other. This principle is captured by the small-world network topology, which have been shown to appear in many domains, including within the network of groups on Facebook [15].

Watts-Strogatz Model

Watts and Strogatz proposed their model for the generation of small-world graphs in [7]. The model is parameterised by the number of nodes n, the mean node degree k and the rewiring probability p, and generates an undirected graph with $\frac{nk}{2}$ edges using the following algorithm.

1. Generate a graph with n nodes each connected to k neighbours i.e. connected to $\frac{k}{2}$ nodes on each side: this is a *regular ring lattice*. More precisely, there exists an edge (i, j) between nodes i and j, if and only if condition 2.7 is met.

$$0 < |i - j| \mod \left(n - 1 - \frac{k}{2}\right) \leq \frac{k}{2}. \tag{2.7}$$

2. Then, consider every edge (i, j) and rewire the end connected to node j to another node k ($k \neq i$ as no self-loops) with probability p: the node k is selected uniformly at random.

Figure 2.4 shows the effects of p on the structure of the resulting network. Watts and Strogatz showed that for p in the interval $[0.01, 0.1]$ graphs generated via this procedure exhibited small-world properties. Such small-world properties include a high clustering coefficient and a low average path length. This structure leads to clusters or *cliques* of highly interconnected nodes, that are then connected by "hub" nodes that act as bridges between the clusters.

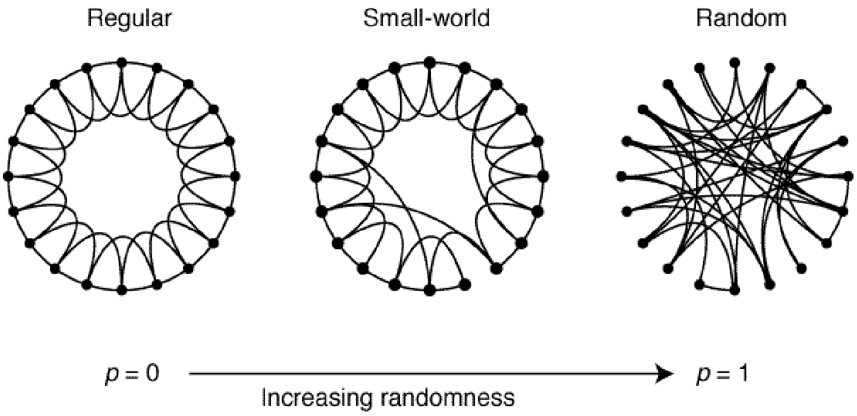


Figure 2.4: An example shown by Watts and Strogatz in [7] of the probabilistic edge rewiring process with different values of p, interpolating between a regular ring lattice and a random network.

2.5.3 Scale-Free Networks

Another widely found property of large, real-world networks is a scale-free power-law distribution of node degrees: this is the underlying property of so called *scale-free* networks. Mathematically, that is the fraction $P(k)$ of nodes in the network having k connections to other nodes (i.e. the probability of a node having degree k) decays as a power law:

$$P(k) \sim k^{-\gamma} \tag{2.8}$$

where typically $2 < \gamma < 3$. The scale-free name arises from the scale invariant nature of the property. This property was first shown to exist in complex networks such as pages of the World Wide Web by Barabási and Albert in [16], and has subsequently been shown to exist in many other domains. This degree distribution leads to a small number of highly connected nodes with large degrees far in excess of the network average: these are often referred to as “hub” nodes. In the fields of social network analysis and social dynamics, these nodes often represent so called “influencers”, and leads to the more colloquial name of influencer networks for scale-free networks. Such nodes often play a pivotal role in the dynamics of the network, due to their high connectedness.

Barabási-Albert Model

Barabási and Albert also presented in [16] an algorithm for generating scale-free networks. The key mechanism of their generative model was *preferential attachment*, a key feature of many real-world networks including online social networks. Intuitively this mechanism ensures that the “the rich get richer”: the more connected a node is in the existing network, the more likely it is that new nodes will connect to it as they are added and the network grows. The model is parameterised by the number of nodes n and the number of edges m which we add to a new node being introduced to the network. The graph generation algorithm proposed by Barabási and Albert is described below.

1. Start with a connected network with m nodes.

2. Add the remaining $n - m$ nodes one at time, connecting each with m nodes in the existing network using the preferential attachment mechanism: the probability Π that

a new node will be connected to an existing node i depends on the prior connectivity k_i of node i:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.9}$$

It was later shown by Albert and Barabási that scale-free graphs generated using this model exhibit small average path lengths, particularly when compared to random Erdős-Rényi graphs [17]. Additionally they showed that BA graphs have a larger clustering coefficient than comparably sized Erdős-Rényi graphs, but a smaller clustering coefficient than small-world networks generated using the Watts-Strogatz model.

Chapter 3

Related Work

In this chapter we present a number of studies into the use of online social networks for news consumption that motivated and shaped our model and experiments. We also present two datasets that we will use as part of our investigation.

3.1 Real World Factors Affecting Information Diffusion

The field of social network analysis has grown significantly in recent years, with many studies and research focussing on the dynamics of online social media. As the pervasiveness of online social media increases the way in which individuals communicate, become informed and adopt opinions also changes. As we discussed in section 1.1, this is because social media results in a disintermediation of news consumption, allowing information to spread without verification. As a result understanding how users make use of and interact with social media is key to understanding its effects on users beliefs. We now introduce the key findings from a number of studies into the real world use of social media.

3.1.1 Increase in use of Social Media for News Consumption

Online services such as Facebook and Twitter have changed the way individuals communicate and absorb information, and their prevalence is only growing more significant. An annual report into digital news consumption patterns [18] showed that in both the UK and the USA, between 2016 and 2019 there was a decline in the use of traditional media (print, radio and TV) for news consumption and a notable increase in the use of smartphones for the same purpose. Additionally, the report in [18, p. 15] showed that in the UK 35% of individuals use social media as their primary source of news, whilst in the USA the figure is even higher at 43%.

3.1.2 Selective Exposure, User Polarisation and Echo Chambers

To better understand the implications of users consuming news through social media sites such as Facebook, Vicario et al. have conducted many studies into the consumption patterns of users on Facebook [5, 6, 19].

In [6], Vicario et al. show how many users are more likely to accept, consume and interact with content that is compatible with their pre-existing beliefs, and will ignore other dissenting information. Such behaviour is described in psychology by the tendency of individuals to exhibit *confirmation bias* i.e. to want to reinforce their own world view. Over time this leads to individuals in a social network becoming even more polarised.

Further research in [20, 21] shows how selective exposure and user polarisation foster the formation of *homogenous clusters* or *echo chambers* online. Again through a quantitative analysis of Facebook data, they show that the level of interaction of a particular user to a certain type of online content is positively correlated to the number of friends that also consume similar content. Such *homophily*, when combined with other behaviours such as selective exposure fosters the formation of like-minded communities online, referred to as *echo chambers*. This terminology comes from the observation that these communities further reinforce the already established views of their members, due to the very fact that everyone has very similar opinions that are continuously reverberated internally.

3.1.3 Example Model for Selective Exposure

Having discussed the broad work of Vicario et al. and others, we now present a more detailed description of a model for selective exposure presented in [21]. In this work, a large quantitative analysis is undertaken to better understand how individuals consume news on Facebook through interaction with Facebook pages. The study looks at 920 news sources and 376 million users, and observes that users focus on a small subset of pages as news sources, leading to a "sharp community structure". To reinforce these findings, Vicario et al. present a model that reproduces these observations by modelling selective exposure. This selective exposure model is built on top of the Bounded Confidence Model (BCM) presented by Deffuant et al. in [22], which we introduce first.

Bounded Confidence Model

The BCM introduced in [22] describes how the opinion's of individual agents in a network change over time due to interaction with each other. The model considers N agents, each with a continuous opinion x_i for $1 \leq i \leq N$. Two agents with opinions x and x' will only interact if the magnitude of the difference of their opinion's is within a threshold d, i.e. $|x - x'| < d$. If this condition holds the two agents interact, and they adjust their respective opinion as a result of the interaction according to equation 3.1. In equation 3.1 the term μ is simply the convergence parameter.

$$\begin{aligned} x &= x + \mu \cdot (x' - x) \\ x' &= x' + \mu \cdot (x - x') \end{aligned} \tag{3.1}$$

The threshold condition is included to account for real world behaviour such as that discussed in section 3.1.2. Deffuant et al. describe the threshold d as the "openness to discussion" amongst agents in the network. They show that a large openness to discussion results in the opinions converging towards an average across the whole population, whilst small openness to discussion thresholds results in several homogenous clusters each with differing opinions. This behaviour is identical to that observed in real world analysis presented in section 3.1.2.

Selective Exposure Model

The model presented in [21] is a modified version of the BCM discussed above, focussing on the relationship and resulting interactions between news sources and individuals on Facebook, with the aim of capturing the behaviour that individuals "interact" more frequently with pages that share news more compatible with their own beliefs. In this context, the term interact refers to the action of a user "liking" a Facebook page.

The core model entities are pages $p \in P$ representing Facebook pages as sources of news and information, and Facebook users $u \in I$. Collectively these two sets of entities correspond to the set of agents described in the BCM. A real number $c_p \in [0, 1]$ is assigned to each page $p \in P$, representing the "editorial line" (i.e. opinion) of the page. Additionally each user $u \in I$ has an initial opinion modelled as a real number $\theta_u \in [0, 1]$. Both page opinions c_p and user opinions θ_u are uniformly distributed.

The model dynamics then focus on the distance between the opinion of a user u and the editorial line of a page p, i.e. $|c_p - \theta_u|$. The concept of confirmation bias is then modelled by assuming that if user u "likes" a page p and the *opinion distance* $|c_p - \theta_u|$ is less than a parameterised threshold Δ, the opinion of user u will converge towards the editorial line of the page p with which it interacted. The convergence of user opinion occurs according to equation 3.1 from the BCM. Specifically for this application, the convergence will occur according to equation 3.2 which makes use of the same notation. θ'_u denotes the new adjusted opinion for user u, whilst θ_u denotes the original opinion prior to adjustment. c_p denotes the editorial line of the page, and again μ denotes the convergence parameter.

$$\theta'_u = \theta_u + \mu \cdot (c_p - \theta_u) = \theta_u + \mu c_p - \mu \theta_u = (1 - \mu)\theta_u + \mu c_p \tag{3.2}$$

Vicario et al. augment the model further by introducing for each user an activity coefficient a_u that represents the number of pages a certain user can visit. As a result, the final opinion of a user u will be calculated by the average of the editorial lines of all the pages that the user "likes". Vicario et al. denote by Ω the set of pages that match the preference of user u: that is, the set of pages p for which $|c_p - \theta_u| < \Delta$. With this notation they describe the average opinion for user u denoted by $\bar{\theta}_u$ according to equation 3.3.

$$\bar{\theta}_u = (1 - \mu)\bar{\theta}_u + \mu |\Omega|^{-1} \sum_{p \in \Omega} c_p = |\Omega|^{-1} \sum_{p \in \Omega} c_p \tag{3.3}$$

With this, we can now fully describe the dynamics of the model for selective exposure presented in [21]. A user u randomly selects a subset of pages from P with which to interact i.e. "like", however the interaction only occurs if $|c_p - \theta_u| < \Delta$. Upon such an interaction, the opinion of user u will be adjusted according to equation 3.2: this adjustment of opinion sees the opinion of u converge towards the average of the editorial lines of the pages with which it interacts. The threshold condition means that u should only interact with pages that are already compatible with the beliefs of u, and as a result u's opinion is reinforced. On average over many simulations, the average opinion of u is given by equation 3.3.

This model provides an example for how real world observations such as selective exposure and confirmation bias can be modelled. It also serves as a basis to allow for the discussion of how our work and model relates to that of Vicario et al. which will be presented in chapter 4.

3.2 SNAP Datasets

As well as generating our own artificial social networks for experimentation purposes, we will also make use of real social networks from the SNAP project. This will provide a closer representation of how individuals are connected across the internet through the use of social media platforms such as Twitter and Facebook, and will better allow us to relate our findings back to the real world.

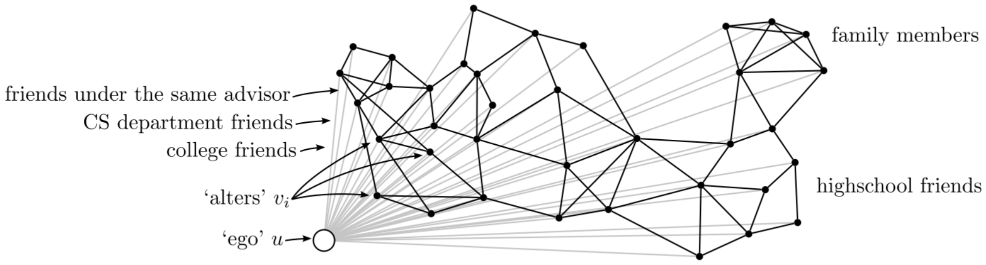


Figure 3.1: An example of an ego-network as shown by Leskovec and Mcauley in [24, p. 2].

The Stanford Network Analysis Platform (SNAP) [23] is a general purpose network analysis and graph mining library. As part of the platform, a number of large network datasets are provided. As part of this project, we will make use of two of these datasets, which we will describe in sections 3.2.2 and 3.2.3. First, we introduce the notion of an *ego-network* to allow us to more accurately describe the nature of the datasets in question.

3.2.1 Ego-Networks

The SNAP datasets used as part of this project were both collated using a discovery process presented by Leskovec and Mcauley in [24]. The datasets are presented as *ego-networks*, that is the network of connections between the "friends" of a particular node.

To explain this notion, let us consider a single node u in an online social network such as Facebook. We then consider each of the nodes v_i that are connected to u, and the subsequent connections between each of the v_i nodes: the node u is called the *ego* of the network as the network revolves around the interconnectedness of u's friends, whilst the v_i nodes are called *alters*. The ego-network associated with node u will be a subset of the entire online Facebook social network, and provides us with a realistic and manageable social network with properties that should be largely consistent and applicable to the network as a whole.

3.2.2 SNAP Facebook Social Circles

The SNAP Facebook social circles dataset consists of a number of anonymised "friends lists" collected from survey participants, that was then used to construct an ego-network relevant to the participating user. The dataset is presented as a collection of 10 smaller networks that can be combined into a single connected example of a Facebook network with over 4000 nodes. Figure 3.2 shows a visualisation of the complete dataset, which clearly shows the high clustering and abundance of "hub" nodes within the network. Table 3.1 shows a summary of relevant dataset statistics for our investigation, compiled from those provided by SNAP and through our own analysis. Crucially, we can see that the network has a relatively high clustering coefficient and a low average path length, the key properties of the small-world network topology discussed in section 2.5.2. This further confirms the validity of our claim that this dataset will provide a more accurate example of real-world social network structure when running our information cascades.

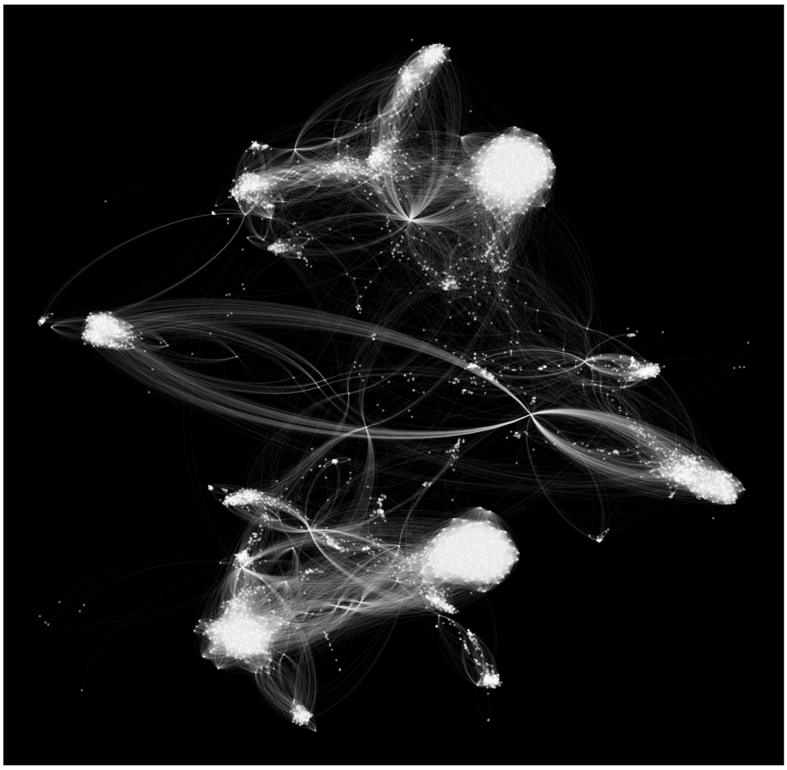


Figure 3.2: A visualisation of the SNAP Facebook social circles dataset in [23] as generated by Sims in [25].

Dataset Statistics	
Nodes	4039
Edges	88234
Average Degree	43.6910
Average Clustering Coefficient	0.6055
Transitivity	0.5192
Average Shortest Path	3.6925
Diameter	8

Table 3.1: A summary of relevant statistics for the combined SNAP Facebook social circles network dataset, compiled from those provided by SNAP as well as through our own analysis.

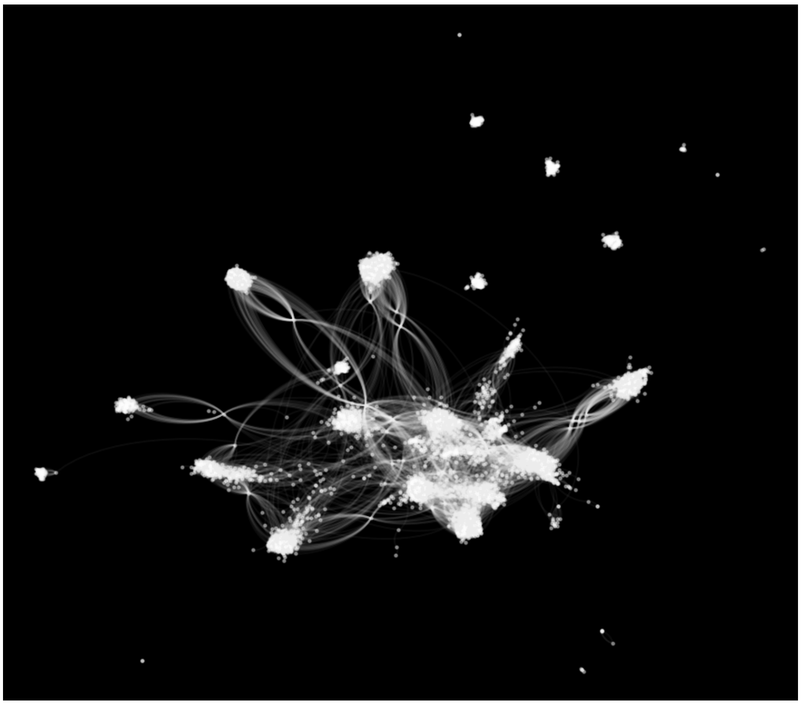


Figure 3.3: A visualisation of our subset of the SNAP Twitter social circles dataset in [23] as generated by the procedure in [25].

3.2.3 SNAP Twitter Social Circles

The second dataset that we will make use of in our investigation consists of anonymised Twitter "lists" collected from public sources. Again, the dataset is presented as a collection of smaller networks that are combined into a larger network, however this combined dataset is too large for our experimental purposes in this investigation. As a result, we have created our own "sub-network" of the full network by combining a subset of the provided networks.

We created this "sub-network" by randomly selecting 25 of the provided 973 small networks and combining them using the ego-network property: the ego of each of the 25 smaller networks is friends with each of the nodes in its network. This resulted in a connected, directed graph that is a subset of the real Twitter network provided in [23]. Table 3.2 summarises the important statistics and properties of the network for our investigation. Most importantly, table 3.2 shows that the network has a lower average clustering coefficient and transitivity (a measure of global clustering) as well as a relatively small average shortest path length: these are key metrics displayed by a scale-free network topology as described in section 2.5.3. The underlying scale-free property can be seen by looking at the degree distribution, as shown in figure 3.4.

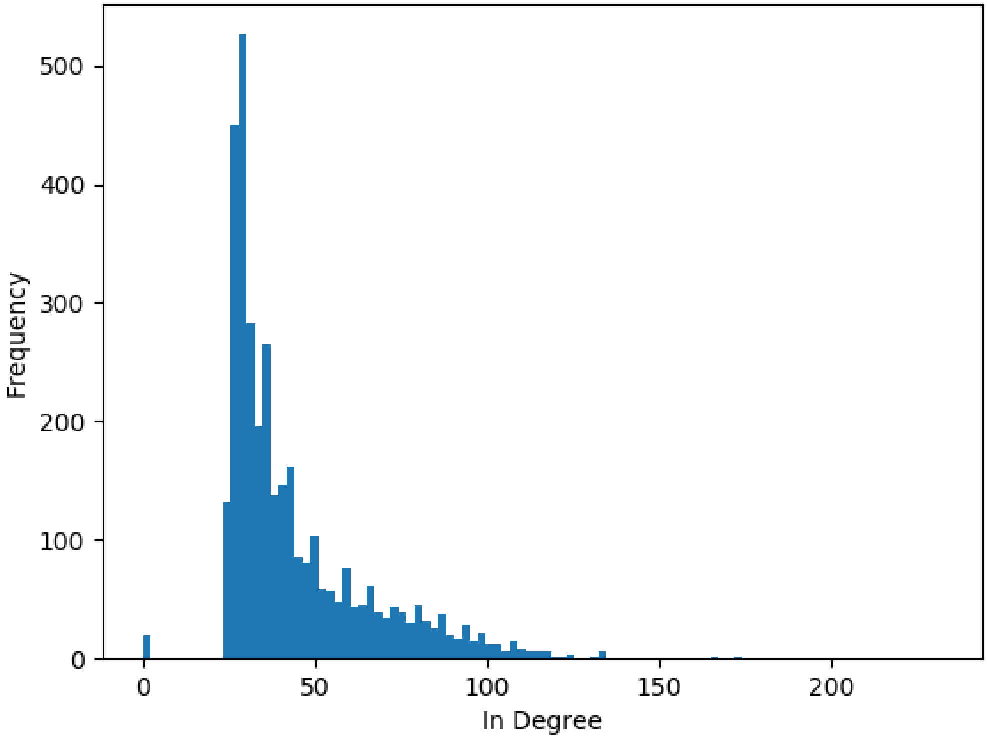


Figure 3.4: Histogram of the in-degree distribution of our subset of the SNAP Twitter network highlighting its approximate scale-free nature.

Dataset Statistics	
Weakly Connected	True
Strongly Connected	False
Nodes	3516
Edges	157425
Average Degree	44.7739
Average Clustering Coefficient	0.3796
Transitivity	0.0114
Average Shortest Path	2.9655

Table 3.2: A summary of relevant statistics for our subset of the SNAP Twitter social circles network dataset collected through our own analysis.

Chapter 4

Model Design

In this chapter we discuss the evolutionary approach we took to designing our model, and how we introduced mechanics one at a time in an attempt to produce the expected model behaviour. At the end of the chapter we evaluate our final model behaviour as well as the behaviour observed after the introduction of each model mechanism.

4.1 Core Model

In this section we introduce the core mechanism of our information cascade model, describe its behaviour and motivate its design and sources of inspiration from the literature.

4.1.1 Structure

In this project we are focussing on how information cascades within social networks are affected by agent state, where such state represents some kind of social product present in society such as political opinion and pre-existing belief systems.

Our model design centres around a threshold condition between entities in the underlying social network of the information cascade, similar to that used in the BCM and selective exposure model discussed in section 3.1.3. Consider that we wish to model an information cascade through a graph $G = (V, E)$ with the set of vertices V representing the set of agents in the social network, and the set of edges E representing the social link between two agents. We assign to each individual $v \in V$ an *opinion* modelled as a real number in the interval $[0, 1]$. We denote the opinion of node $v \in V$ by θ_v for notational consistency with the selective exposure model discussed in section 3.1.3. Figure 4.1 shows a simple example of this. A seed node is chosen, which will be the origin of the information cascade. The modelling of the cascade then proceeds as a discrete time simulation, with the cascade dynamics occurring as follows.

At each discrete time step, an “activated” node (i.e. a node that is already part of the information cascade) has the opportunity to interact and “spread” the information cascade further with each of its (outgoing) neighbours, however the interaction only occurs if a threshold condition is met. The threshold condition for an interaction to occur between two nodes u and v where u is already part of the cascade and $v \in N(u)$ if G is undirected or $v \in N^{out}(u)$ if G is directed, is given by:

$$|\theta_u - \theta_v| < d \tag{4.1}$$

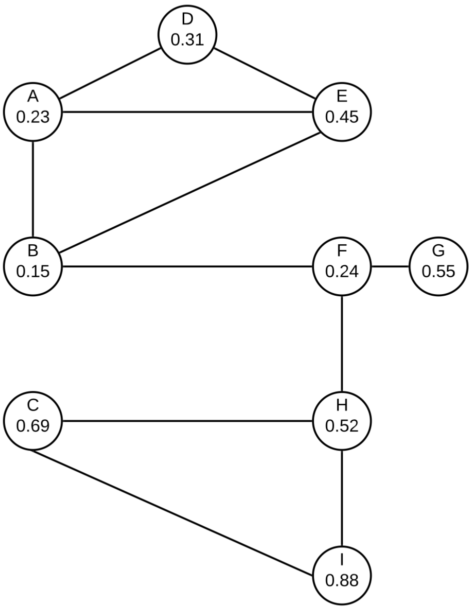


Figure 4.1: An example of a simple graph representing a social network, in which each node has been assigned as opinion in the interval $[0, 1]$.

where d is the parameterised *opinion threshold*.

4.1.2 Motivation, Influence and Link to Literature

The motivation for this simplistic initial model design revolving around the assignment of an "opinion" to each node is that in reality belief systems exist on a continuous rather than discrete or binary scale. By modelling beliefs in the interval $[0, 1]$ we are able to capture the sense of a true spectrum, where the closer the numerical value to 0 or 1 indicates a highly *polarised* world view whilst 0.5 suggests a more typical "balanced" view. Additionally it makes sense to assume that two individual's ideas must normally be sufficiently similar to even warrant a potential interaction or "discussion". We see these mechanics as a basis for adding additional assumptions/features if and when we require them to obtain the correct model behaviour.

The core of our model, the *opinion distance* threshold condition is similar to the mechanics of models presented in the previously discussed background such as [21, 22], however we emphasise that we are using these similar mechanics to study different problems. The selective exposure model presented in [21] is primarily focussed on the interaction between news sources and agents, and how through individuals choosing to interact with certain confirmatory news sources their state is adjusted to further reinforce their beliefs. Similarly the bounded confidence model presented in [22] is interested in analysing how interactions affect agent state, or social products using our terminology.

Meanwhile, our work looks to focus on how agents themselves affect information flow due to their pre-existing social products such as political beliefs. We assume that once the cascade begins all agent state remains constant: in this regard we are approaching

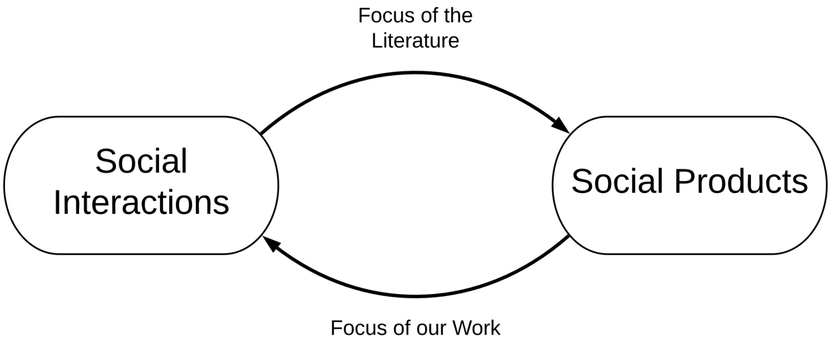


Figure 4.2: An illustration of how the two core concepts in information cascade models are related.

a very similar problem to those studied in the literature from the opposite angle. Our work focusses on the *social* impact on information flow, whilst Vicario et al. have focussed on modelling the initial information spread and consumption. This distinction is shown graphically in figure 4.2.

4.1.3 Model Dynamics Example

To reinforce our models behaviour, we now walk-through a simple example of an information cascade step-by-step.

We start from the network shown in figure 4.1, in which each node has already been assigned an opinion. We then refer to figure 4.3 which shows the different discrete time steps of the cascade: for this example we take the opinion threshold to be 0.1. When $t = 0$, the only node in the cascade is the seed node which was randomly chosen to be node B depicted in yellow.

At $t = 1$, node B is able to further the cascade to include each of its neighbours if the opinion threshold condition is met. Node B is able to spread the cascade to nodes A and F, however $|\theta_B - \theta_E| > 0.1$ and so node E is not added to the cascade as shown by its grey colouration. At $t = 2$, node A influences node D as the threshold condition is met, however both of node F's neighbours are not influenced. As a result, the remaining nodes in the network are unreachable and so the cascade stops after $t = 3$ iterations.

4.2 Pre-Cascade BCM Phase

After our initial model implementation we introduced a new mechanic to the model where a separate phase is run prior to the cascade simulation.

4.2.1 Motivation

In the initial model discussed above, the node opinions are assigned according to the parameterised distribution: for basic analysis of our model behaviour, we assume this distribution to be uniform however other distributions will be considered as part of our experimentation presented in chapter 6. Whilst this allows the model to capture the fact that opinions of individuals across a network as a whole are widely spread, it does not

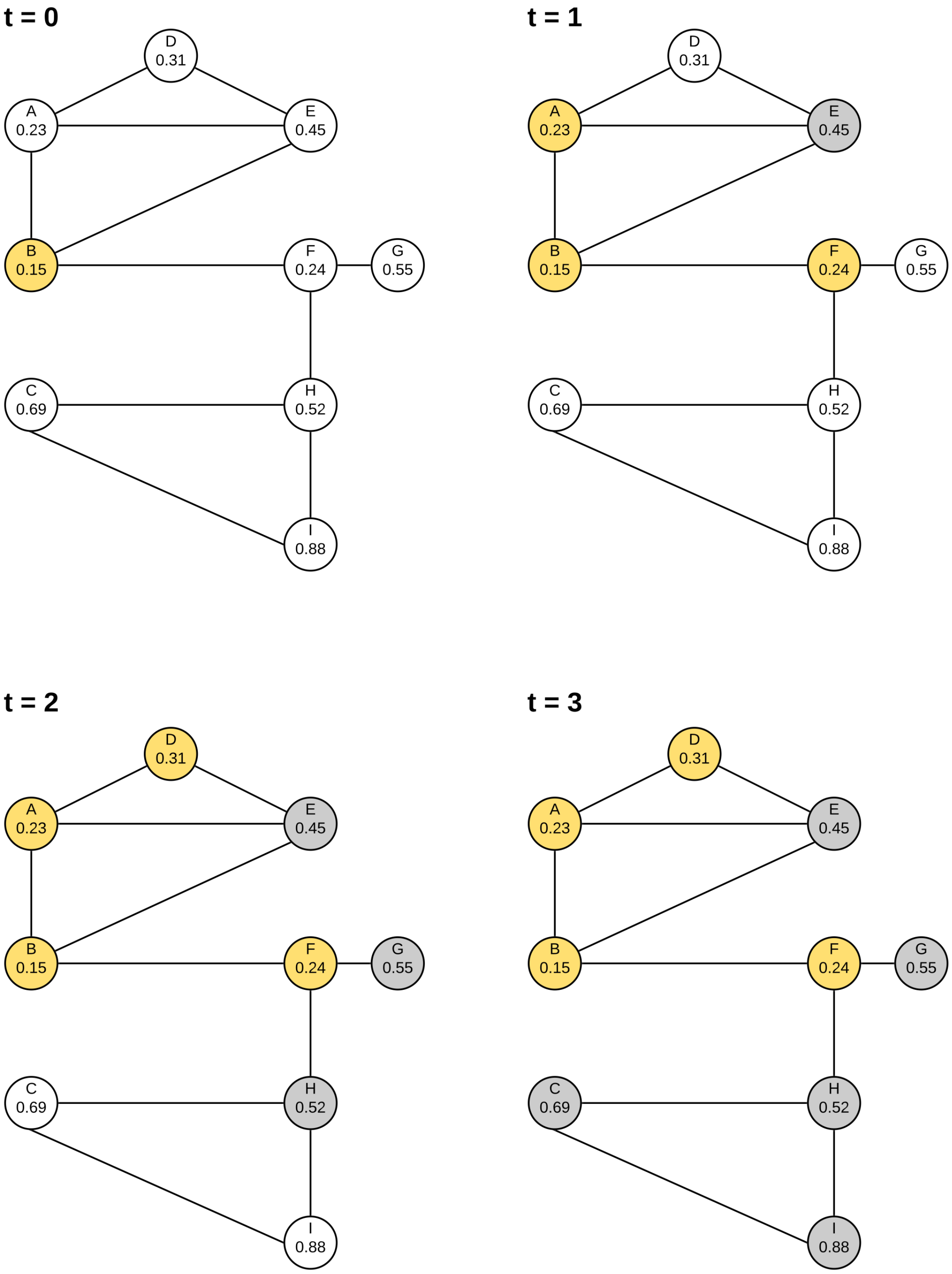


Figure 4.3: An example of a simple information cascade using our model, in which we take the opinion threshold to be 0.1.

accurately reflect the fact that pairs of nodes in a network that are already connected (i.e. that are already "friends") have interacted in the past and this might have affected their opinions.

To try and capture some of this behaviour in the network prior to running our information cascade, we introduced a preliminary phase in the model setup which incorporates a form of the bounded confidence model (BCM) mechanics presented in [22] and previously discussed in section 3.1.3. This mechanism allows the opinions of neighbouring nodes to converge, simulating how the outcomes of *previous* cascades might have affected agent state and simultaneously allowing our model to incorporate the notion that individuals tend to be friendly with those that have similar views.

4.3 Stochastic Element

The final element of the model that we added is a probabilistic element to the actual cascade condition. Prior to this, if the opinion threshold condition was met then the interaction would always be regarded as having occurred and the neighbouring node would be added to the cascade.

4.3.1 Motivation

In attempting to model an information cascade, we wish to capture the phenomenon of individuals adopting ideas as a result of *social* interactions: this is modelled by neighbouring nodes being added to the cascade if the threshold condition is met. However by allowing the cascade to occur in a deterministic manner, we are unable to capture the idea of two individuals having a social discussion that can result in multiple outcomes.

This is the motivation behind the addition of some element of randomness to the cascade mechanism. The opinion threshold condition still exists as we assume that the opinions of two neighbouring nodes must be sufficiently similar to allow for a discussion to occur: the term "openness to discussion" was used by Deffuant et al. to describe this notion in [22]. If this condition holds and the two nodes are "open to discussion" then the probability of the interaction resulting in "success" (which we define to mean that the cascade has been furthered) is determined by true distance between the opinions of the neighbouring nodes.

As an example consider we are modelling the information cascade in a graph $G = (V, E)$ with opinion threshold d and neighbouring nodes $u, v \in V$ where u is already in the cascade. Given that $|\theta_u - \theta_v| < d$, the probability that the information cascade continues from u to v is given by equation 4.2.

$$p = 1 - \frac{|\theta_u - \theta_v|}{d} \tag{4.2}$$

Clearly this captures the behaviour that the smaller the distance between the two interacting nodes opinions, the higher the probability that the interaction is successful: this is intuitive as the closer the two individual's opinions are the more likely the neighbouring node is to adopt the information being conveyed.

4.4 Evaluation

In this section we analyse and evaluate the behaviour of our model. We briefly discuss the behaviour of the model after the introduction of each new feature and describe how this motivated the addition of subsequent features, before analysing the final model behaviour.

4.4.1 Process

For the purposes of this initial model behaviour evaluation we used the SNAP Facebook example network as discussed in section 3.2.2. We chose this network for our initial evaluation due to its typical small-world properties and modest size. We also used uniformly distributed node opinions, and the seed node was chosen uniformly from all nodes in the graph. We then proceed with the information cascade allowing it to run to completion i.e. until no more nodes can be added to the cascade. Upon completion we form the cascade graph from all the edges added during each iteration, which we use to perform our analysis of this particular cascade. To reduce the effects of fluctuations in the stochastic process we perform 100 simulations to allow us to average the results.

4.4.2 Metrics

There are a number of graph metrics that can be helpful when quantitatively analysing and comparing graphs. For evaluating our model behaviour and throughout our investigation, we will focus on two metrics when comparing cascades.

Mean Edge Homogeneity

The metric of mean edge homogeneity was introduced by Vicario et al. in [6] during their analysis of the way in which individuals interact with news sources online. It reflects the average similarity of two nodes views in relation to the scale of possible opinions possible.

As we have discussed each node in the underlying network of our model is assigned a real-valued opinion in the interval $[0, 1]$. This represents an overall modelling of the nodes beliefs or world view, on a scale where 0 and 1 are deemed more polarised and "extreme" views whilst 0.5 suggests a more typical, average opinion. Then using the opinion, we additionally define for each node its *polarisation*. For a node $u \in V$ of a graph $G = (V, E)$ its polarisation σ_u is defined by:

$$\sigma_u = 2\theta_u - 1 \tag{4.3}$$

where θ_u is its opinion. Clearly as $0 \leq \theta_u \leq 1$ then $-1 \leq \sigma_u \leq 1$. The polarisation of a node reflects how polarised its opinion is from the average. From the polarisation, we can then define the *edge homogeneity*: for an edge $(u, v) \in E$, the edge homogeneity is defined as:

$$\sigma_{uv} = \sigma_u \times \sigma_v \tag{4.4}$$

where σ_u and σ_v are the polarisations of nodes u and v respectively. Given that $-1 \leq \sigma_u, \sigma_v \leq 1$ we have that $-1 \leq \sigma_{uv} \leq 1$. The edge homogeneity measures how similar the polarisation's of two connected nodes are and then any edge with positive edge homogeneity is defined to be a *homogenous* link i.e. the polarisation of the two nodes opinions are more positively correlated than negatively correlated [6]. When considering whole cascades particularly within larger networks it is more useful to consider the *mean edge homogeneity* of the cascade, that is the mean average of the edge homogeneity of every cascade edge.

The mean edge homogeneity is a random variable over the sample space of all possible information cascades within a given network. As we are repeating each experiment 100 times to avoid stochastic fluctuations, we will be able to estimate the probability density of the mean edge homogeneity for a given network and set of model parameters. If we observe high probability densities for positive mean edge homogeneities, we can conclude that the majority of links in an information cascade occur between like-minded individuals.

We choose to include mean edge homogeneity as a key metric in our investigation to allow us to easily compare our findings with those made it related experiments in the literature, such as those presented in [6]. Whilst the metric was introduced for a slightly different investigation into how users interact with news sources online, the principle of quantitatively measuring the similarity of links in a network is applicable to our use case.

Cascade Depth

The other key metric we will consider during our analysis is the cascade depth. We can consider the cascade graph as a spanning tree of nodes in the cascade, with the root node of the tree being the seed node of the information cascade. The depth of the cascade tree is then the longest path from the root node to a leaf, signifying the furthest the "idea" has been able to spread from the seed. Due to the way we grow the cascade, this depth is equal to the number of iterations required for the cascade to complete. This metric will be particularly useful when we investigate the role of other factors such as network topologies in the information cascade.

4.4.3 Initial Model Behaviour

We start by evaluating the initial model with just the opinion threshold condition as described above in section 4.1. We vary this threshold in the range $[0, 1]$ and for each model parameter we perform 100 model simulations and calculate the mean edge homogeneity. We find that as the opinion threshold goes to 1, the average mean edge homogeneity goes to 0. Therefore we focus on the interval $[0, 0.05]$ in order to maximise mean edge homogeneity. For a more detailed discussion of model behaviour in the full threshold interval $[0, 1]$ see appendix A.

Figure 4.4 shows an example information cascade with the opinion threshold set to 0.05. We can see that the cascade resulted in two identifiable clusters based on the edge homogeneity: qualitatively this is the desired behaviour we are aiming for based on real-world observations in the literature as was discussed in chapter 3. To quantitatively assess our model's behaviour, we estimate the probability density function for the mean edge homogeneity from our sampled results at intervals in the range $[0, 0.05]$ as shown by figure 4.5. We truncate the plots to the positive domain as we observed no instances of negative mean edge homogeneity from our sampling. We see that for smaller opinion thresholds we observe a more uniform probability density in the range $[0, 1]$, and as the threshold increases the standard deviation of observed homogeneities decreases and the mean approaches 0: again see appendix A for further discussion. We see expected model behaviour in that we observe no 0 or negative homogeneities suggesting that most cascade edges are homogenous, however the lack of peak in the higher homogeneity range suggests model adjustments can be made.

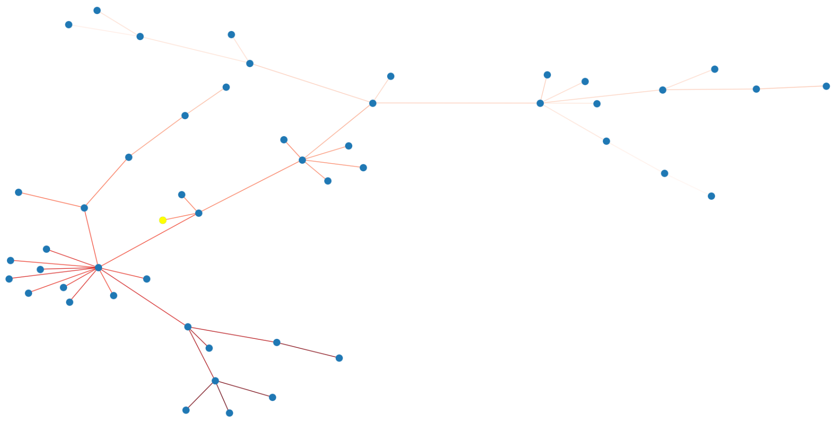


Figure 4.4: An example cascade generated by our simulation of the SNAP Facebook network with opinion threshold 0.05. The seed node was chosen randomly and is highlighted in yellow. The saturation of the edge colour describes the edge homogeneity: darker edges have higher homogeneity whilst lighter edges have lower homogeneity.

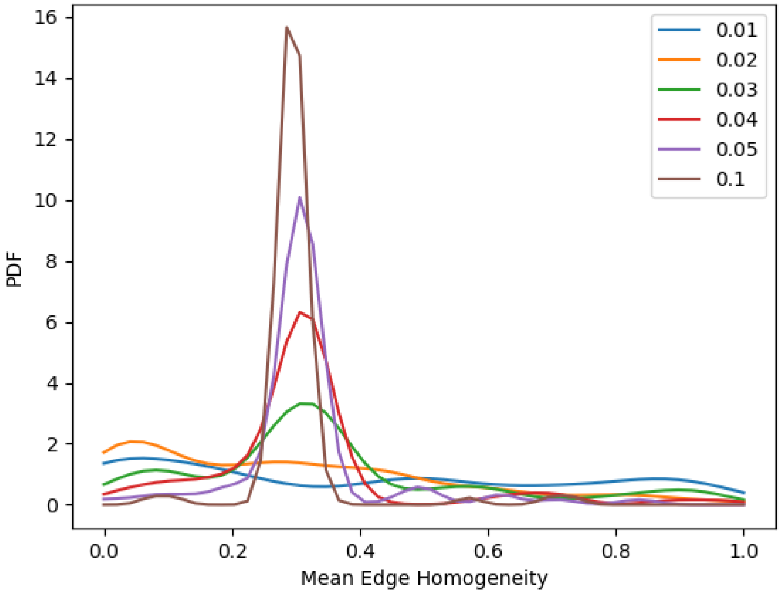


Figure 4.5: PDF plots for the mean edge homogeneity of cascades with opinion thresholds in the range $[0, 0.1]$ for the initial model evaluated on the SNAP Facebook network.

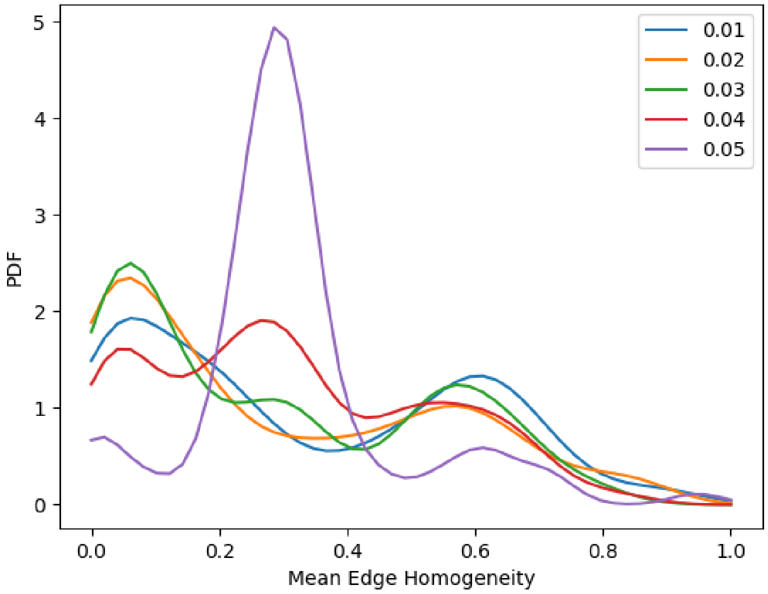


Figure 4.6: PDF plots for the mean edge homogeneity of cascades with opinion thresholds in the range $[0, 0.05]$ for our model with the addition of BCM-inspired convergence phase.

4.4.4 Addition of BCM Convergence Phase

The introduction of a BCM-inspired convergence phase similar to that discussed in section 3.1.3 allows the opinions of neighbouring nodes to converge, better modelling the real-world tendency for individuals to have to closer opinions to their neighbours. This mechanism is parameterised by the *openness threshold* and the convergence parameter μ: the openness threshold plays a similar role to the opinion threshold, however its value can be distinct as it is used in an earlier phase of the simulation. These values function similarly to hyperparameters in machine learning optimisation, and so we performed a parameter search by fixing the opinion threshold at 0.01 and varying the openness threshold and μ in the interval $[0, 0.5]$. We found that an openness threshold of 0.1 and convergence parameter of 0.25 were optimal, allowing for sufficient convergence of opinion prior to the cascade resulting in improved model behavior.

Having found sensible values for these parameters, we then again plotted the mean edge homogeneity as a function of the opinion threshold as shown in figure 4.6. The sharper peak as the threshold increases is expected due to the high clustering of the underlying network: again see appendix A for further discussion. However notably for opinion threshold less than 0.05, we observe an almost bimodal distribution with peaks around 0.1 and 0.6. We attribute this observation to the fact that BCM phase allows neighbouring node opinions to converge slightly, meaning that on average edge homogeneity is also higher.

We also consider the effects of the opinion threshold on the average cascade depth. As the threshold increases in the interval $[0, 0.1]$ the cascade depth also increases as the increased *openness to discussion* between nodes allows the cascade to reach further. Then as the threshold increases further past 0.1 we observe that the cascade depth begins to fall. We attribute this behaviour to the strong clustering of the underlying network: once

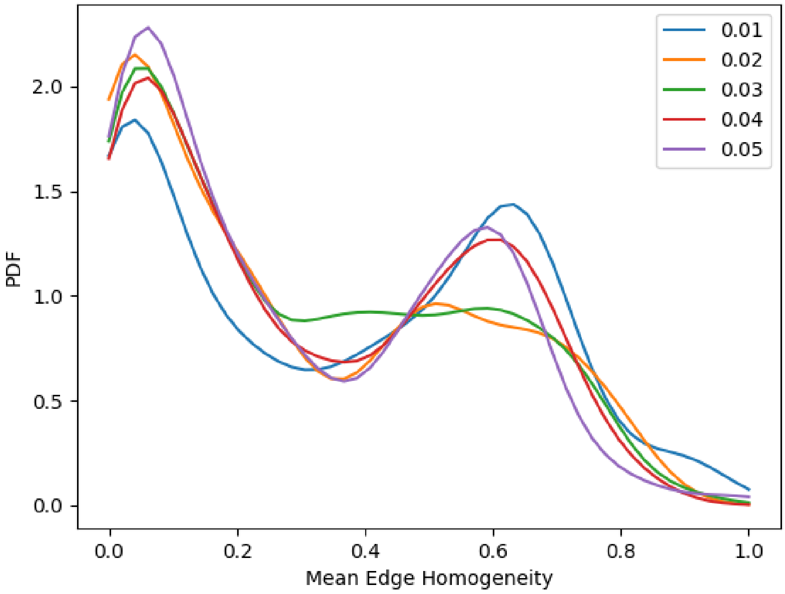


Figure 4.7: PDF plots for the mean edge homogeneity of cascades with opinion thresholds in the range $[0, 0.05]$ for our model with the addition of BCM-inspired convergence phase and the randomness of the cascade condition.

the threshold is sufficiently large the high clustering means that the cascade can rapidly reach a highly connected *hub* node in the network, allowing the cascade to spread across the entire network.

4.4.5 Effects of Randomness

Finally we evaluate the model behaviour with the addition of the stochastic element discussed above in section 4.3 on top of the BCM convergence phase. Figure 4.7 shows the same density plot for this series of simulations. We again see that the smaller the opinion threshold the more closely the distribution appears to be bimodal, however the added nature of chance suppresses the behaviour of the underlying uniform distribution as the threshold increases.

Average Cascade Depth		
Opinion Threshold	Initial Model	Stochastic/BCM
0.01	2.30	2.07
0.02	7.85	2.73
0.03	21.81	3.49
0.04	24.55	4.23
0.05	24.94	4.97
0.1	22.17	7.18

Table 4.1: A summary of the average cascade depths across different versions of our model design.

We also see a notable change in the average cascade depth. Table 4.1 summarises the average cascade depth across different opinion thresholds for both our initial model with the simple threshold condition and our final model. We can see that the average depth is substantially smaller after the introduction of an element of chance: this makes sense as clearly the stochastic mechanic gives the cascade a chance to stop at each new node.

It is also worth highlighting the interesting jump in cascade depth for the initial model between opinion threshold 0.02 and 0.03. We expect that this behaviour is heavily tied to the moderate clustering in the underlying social network structure: given the right topology and initial seed node a small increase in threshold can cause a large increase in possible cascade depth. Of course the effects of the choice of seed node should be cancelled out by the stochastic nature of its selection, and so it is highly probable that the main cause of this spike is network structure. Intuitively this also explains the decline in cascade depth after the peak, as once the threshold becomes sufficiently high the density of the network allows information to diffuse faster and so reduces the depth of the cascade.

Broadly we are happy with the behaviour of our final model. Qualitatively we can observe distinct branches of the cascade graph representing interactions within homogenous clusters of the underlying network. Quantitatively we observe no instances of negative mean edge homogeneity throughout our simulations meaning we observe high probability of homogenous links in information cascades, portraying similar behaviour to that observed in the literature.

4.4.6 Strengths

We now summarise the strengths and weaknesses of our model design in relation to the objectives previously set out.

- By focussing our design on the way that information spreads *socially* as a direct result of social interactions, we have been able to cover new ground not focussed on by the literature.
- Our final model exhibits intended behaviour, simulating how individual's beliefs affect information flow in a highly clustered underlying network.
- The iterative approach to model development allowed us to keep our model as simple as possible, by quantitatively assessing its behaviour after the introduction of each new mechanism and allowing us to consider what new mechanisms will give the desired behaviour.
- We have essentially developed two similar but distinct models: one with a deterministic binary spreading system, and another in which chance is introduced.

4.4.7 Weaknesses

- Our model is unable to reproduce the exact behaviour of extreme polarisation observed in real-world cascades as presented by Vicario et al. in [6]. This could be limited by the assumptions made regarding the underlying network such as node opinion and network topology. We hope to test this hypothesis in chapter 6.
- We did not have access to a dataset with sufficient features to allow us to truly map real-world user opinions to numerical values in our model. This restriction is what

motivated us to the quite strong assumption regarding the opinion distribution of nodes within the underlying network. We discuss this further in section 7.2.2

Chapter 5

Implementation

In this section we outline an overview of the implementation details for our model, and the subsequent simulations and experiments. We chose to use Python 3 for our model implementation and the various scripts for running simulations, due to its wide range of available data science libraries. Section 5.2 outlines the main Python libraries we made use of.

5.1 Overview

Figure 5.1 shows a diagram of the pipeline of the execution of a model simulation within our implementation. The core part of our pipeline is contained within the execution of a single Python program as demonstrated by the solid connector arrows, whilst the visualisation aspects function separately and are signified by the dashed connector arrows. This has been done intentionally as visualisation is slow and unnecessary for every simulation iteration: therefore by removing it from the core program we improve the speed and ease at which large batches of simulations can be performed.

The model parameters are supplied to the simulation via a Python dictionary, which are used to setup the simulation. The underlying graph representing the social network being simulated are created or loaded by the NetworkX Python package accordingly. Once the simulation is complete a number of metrics relating to the cascade simulation are calculated, and the results are written to a CSV file.

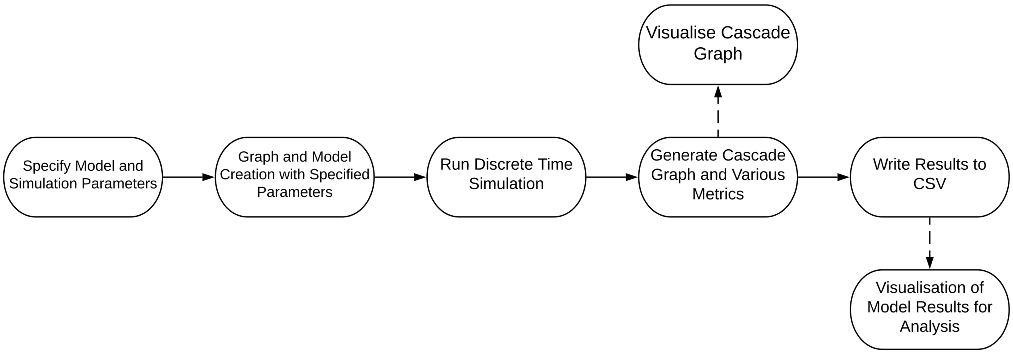

Visualise Cascade Graph
Specify Model and Simulation Parameters
Graph and Model Creation with Specified Parameters
Run Discrete Time Simulation
Generate Cascade Graph and Various Metrics
Write Results to CSV
Visualisation of Model Results for Analysis

Figure 5.1: An illustration of the high level structure of the internal pipeline of our model and simulation.

5.2 Libraries

We now briefly discuss and justify the core libraries we made use of throughout the project.

- **NetworkX** [26] is used to create, represent, manipulate and analyse complex mathematical graphs and networks: we use NetworkX to represent all the underlying social networks in our simulations. We chose to use NetworkX over alternative network analysis packages such as iGraph[1] or Graph-Tool[2] due to its excellent documentation allowing for easy learning, and wide range of in-built graph algorithms allowing us to develop fast and focus on model/simulation design. Whilst NetworkX's pure Python implementation means it is noticeably slower than the alternatives (built on top of C), as we are mostly focussing on smaller network simulations (5000 nodes or less) throughout this project this downside is not too limiting.

- **Pandas** [27] provides powerful features for data analysis and manipulation, with excellent tools for reading and writing data to/from different file formats including CSV. This was incredibly useful when accessing results from CSV files for analysis or when creating visualisations. We chose Pandas for this due to its popularity, performance and more powerful features than comparable packages such as NumPy.

- **Matplotlib** [28] is a particularly well-known Python library for generating charts and other data visualisations: it is closely integrated with NetworkX in terms of creating basic graph visualisations, and is the "standard" Python library for plotting charts such as histograms which is an important part of our project.

- **Dash** [29] is a Python framework for building web applications, built on top of the Python Plotly[3] graphing library: this allows for the creation of interactive graph and chart visualisations, tackling one of the core limitations of Matplotlib which has limited support for interactive plots. We did not make extensive use of Dash or Plotly, however we did use it to create a simple web application for displaying interactive graphs via a local web server: this was useful during the analysis of our simulation results.

5.3 Challenges

5.3.1 Result Storage

Like any simulation or modelling based project we were always going to generate a significant amount of numerical data as part of our experiments that would need to be analysed effectively to evaluate our model. For this reason reliable and efficient storage would be important, with a requirement to be able to quickly select and filter data. Early on in the project development we used CSV files to store results due to their simplistic nature, with the expectation to move to a more complex system as the project progressed. However, we found that properly organised directories and CSV files combined with the Pandas Python library offered us a simple and yet reasonably robust and powerful storage/analysis system.

[1] https://igraph.org/python/

[2] https://graph-tool.skewed.de

[3] https://plotly.com

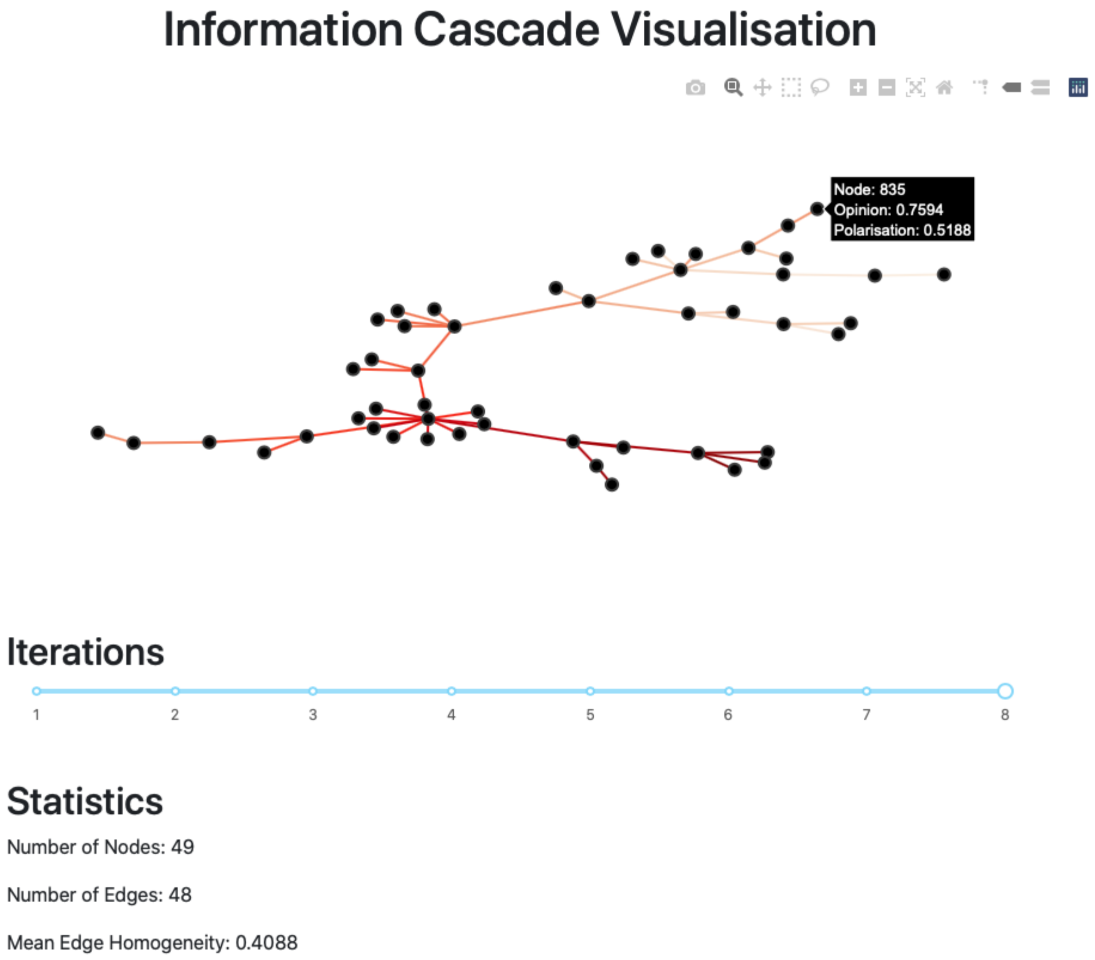
We briefly considered using a remote database for the storage of simulation parameters and subsequent simulation results however we decided that this system would most likely be overly complex for our requirements. Throughout the project development we found that the bulk of our time was spent on designing the model and analysing the resulting cascade data. As this was always intended to be the focus of our project the decision to not spend more time setting up a more complex pipeline proved beneficial as it allowed us to focus on the core project objectives. We discuss in chapter 7 how a better integrated pipeline could be a part of future work.

5.3.2 Visualisation

The biggest challenge we faced with the implementation of the simulation pipeline and model analysis tools was the model visualisation. Whilst NetworkX is a very rich and full-featured Python library for social network analysis its plotting and visualisation tools are quite limited as they rely on Matplotlib. This means that any plots drawn directly using the methods provided by NetworkX are limited to static plots that could be generated by Matplotlib. To better visualise the way in which the information cascade grows at each discrete time step we wanted to develop some form of interactive plot that can display this temporal dimension, as well as other graph metrics on demand.

This was not a key objective of the project and so to simplify the process we sought to find a library or framework to perform this task within Python so that we could ensure smooth compatibility with our NetworkX graph representations. This led us to make use of the Dash Python framework as was introduced above. As we were unfamiliar with this framework during our development we modified code snippets from [30] as an example of how interactive graph and network rendering can be performed using Dash.

We were able to build a simple locally hosted web application for interacting with the information cascade of previously run simulations. Figure 5.2 shows a screenshot of our simple interactive application. When a node or edge is hovered over a summary of relevant information is displayed. The slider allows the user to "scrub" through the discrete steps of the simulation to see how the cascade evolves. The darkness of the edge colour signifies the mean homogeneity: a darker hue represents a edge homogeneity close to 1 whilst a pale colour represents a homogeneity close to 0.



Information Cascade Visualisation

Node: 835
Opinion: 0.7594
Polarisation: 0.5188

Iterations

Statistics

Number of Nodes: 49

Number of Edges: 48

Mean Edge Homogeneity: 0.4088

Figure 5.2: Screenshot of our basic application for allowing interactive information cascade visualisations.

Chapter 6

Experimentation and Analysis

In this chapter we present a number of experiments we conducted to further test our model behaviour, and to additionally look at the roles other factors can have in the way information spreads across social networks online.

6.1 Experimental Process

To ensure all results are comparable when performing simulations to determine model behaviour and as part of experiments, we adopt a standard experimental process.

- Before each simulation all nodes in the network are assigned a real-valued opinion according to the opinion distribution which is a model parameter. All nodes start in an “inactive” state, apart from the seed node which is chosen at random.
- At each discrete time step of the cascade we record the edges (and therefore nodes) that are added to the cascade.
- We allow the cascade to run to completion i.e. until there are no more nodes that could possibly be added to the cascade. This is computationally viable due to the static nature of agent state within our model, and the fact that we are mostly experimenting with reasonably small networks.
- Once the cascade has completed we form the *cascade graph* by collating the edges added during each iteration: this graph is a spanning tree of all the nodes present in the cascade. We can then analyse individual cascades by calculating metrics based on the cascade graph.
- Due to the stochastic nature of each simulation run, for each set of model parameters we perform 100 simulations to allow us to average any results to combat any biases or fluctuations.

To compare the results, we again focus on the cascade depth and mean edge homogeneity as was described previously in section 4.4.2 to evaluate our initial model behaviour.

6.1.1 Factors

Before performing any experiments, we must decide what factors we are considering and the range of these parameters we wish to focus on. For the remainder of this project, we choose to focus on the effects of changing the node opinion distribution and the underlying network topology. We chose to focus on these two factors as they are arguably the most observable and measurable differences that could occur between social networks, and they are likely the most common distinctions between different platforms. For example the directional nature of relationships on networks such as Twitter and Instagram is different to the bidirectional relationships present on Facebook and this influences the network topology.

6.2 Role of Network Topology

6.2.1 Motivation and Setup

Table 6.1 shows a summary of the different network topologies we will use in our experiments. We now briefly motivate the inclusion of each network topology.

- *Small-World* networks exhibit high clustering and low average path lengths. This topology is observed across many different domains including real-world human social networks and the network of groups on Facebook. As described in section 2.5.2 we can generate small-world networks using the Watts-Strogatz generative graph model. We conducted a parameter search to determine optimal parameter values to give an underlying network with high clustering and low path length and diameter: we found that such optimal values for our use case were $N = 5000, K = 50, P = 0.01$.

- *Scale-Free* networks are also observed in real social networks such as Instagram as explained in section 2.5.3. Key properties include low clustering and small average path lengths. Such networks can be generated using the Barabási-Albert model. We chose $N = 5000$ to give a large network and found that $M = 50$ gave us optimal characteristics.

- *Erdős-Rényi* random graphs are uncommon in real-world settings however they provide a good baseline to compare other topologies against. We again chose $N = 5000$ to give a large network size and then set $P = 0.01$ after a parameter search to give minimum clustering and maximum path length.

- The *SNAP Facebook* and *SNAP Twitter* datasets are included as examples of real social networks and are more thoroughly discussed in sections 3.2.2 and 3.2.3 respectively.

We now divert briefly to present our approach for generating directed scale-free graphs.

Directed Scale-Free Topology

All previous network topologies discussed are examples of undirected graphs, apart from the example SNAP Twitter network. To complement this, we wish to experiment with an artificial directed scale-free network to mimic the kind of structure that can be observed in online social networks such as Instagram.

Our implementation library of choice NetworkX provides access to a method for generating directed scale-free graphs based on a process described in [31], however in our testing of this generative model's behaviour we found that the resulting graphs were of extremely low density. Therefore we decided to attempt to devise our own method for the generation of directed scale-free graphs.

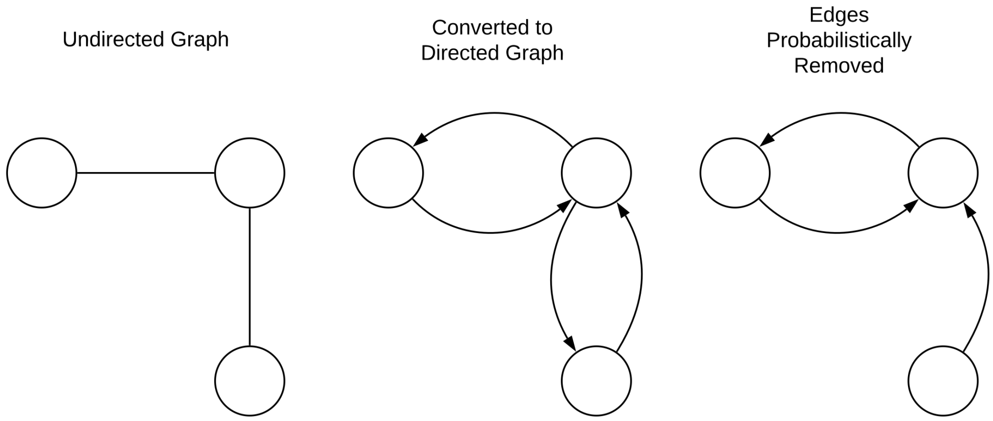


Figure 6.1: A simple visualisation of the steps in our process to generate a directed Barabási-Albert scale-free network.

Our generative model is quite simple. Firstly we generate a directed scale-free network using the standard Barabási-Albert model. Once this network has been generated we convert each undirected edge in the graph to two directed edges, one in each direction. We then probabilistically remove an edge based on the relative in-degrees of the two neighbouring nodes. More precisely let $G = (V, E)$ represent a directed scale-free network generated using the standard Barabási-Albert model, where each undirected edge $(u, v) \in E$ has been replaced with two directed edges $(u, v) \in E$ and $(v, u) \in E$. Then d_u^{in} and d_v^{in} are the in-degrees of the respective neighbouring nodes. Let us assume without loss of generality that $d_u^{in} \geq d_v^{in}$. We say there is a probability of 0.1 that both directed edges remain: this hyperparameter was found through experimentation, and was added to ensure that not every pair of connected nodes has just a one-way relationship. Independently we perform a Bernoulli trial in which we remove the directed edge (u, v) i.e. the edge from the more "popular" node to the less "popular" with the probability shown below: otherwise we remove the directed edge (v, u).

$$p_{remove} = \frac{d_u^{in}}{d_u^{in} + d_v^{in}} \tag{6.1}$$

The reasoning behind this mechanism is to in some sense replicate preferential attachment in which it is more likely that nodes will "follow" other nodes with high degrees. Similarly it is less likely that a high degree node will follow a less "popular" node.

Figure 6.1 shows a simple example of this process. The leftmost network is a simple undirected graph that could form part of a standard scale-free network as generated by the Barabási-Albert model using preferential attachment. We then take each undirected edge and replace it with two directed edges. Finally we remove one of the two directed edges between a pair of nodes based on their relative in-degrees: in figure 6.1 the centre node has two incoming edges whilst the other two nodes only have one therefore meaning it is more likely that we remove the edge from the higher degree node to the lower degree node. Whilst we do not show rigorously that our produced networks follow the scale-free topology, we show in appendix B the degree distributions of an example network generated using this process. These distributions clearly show an approximate scale-free structure in which the in-degree and out-degree distributions follow different power laws.

Topology	Nodes	Edges	Directed	Transitivity	Average Clustering Coefficient	Average Shortest Path	Diameter
Small-World	5000	125000	False	0.714	0.714	4.039	6
Scale-Free	5000	247500	False	0.057	0.057	2.125	3
Random	5000	124952	False	0.010	0.010	2.591	3
SNAP-Facebook	4039	88234	False	0.519	0.606	3.693	8
SNAP-Twitter	3516	157425	True	0.011	0.380	N/A	N/A
Directed Scale-Free	5000	288768	True	0.066	0.030	2.679	4

Table 6.1: A summary of the network topologies we plan to test.

6.2.2 Analysis

Figure 6.2 the average mean edge homogeneity and average cascade depth as functions of the opinion threshold for all the previously described network topologies. For the average mean edge homogeneity we largely see the same trend across all the different networks: this is expected as the opinion threshold increases. For this reason we largely ignore this metric for this experiment.

The most interesting findings occur in the cascade depth, where we see quite different behaviours across different networks. The most striking outlier is that of the Watts-Strogatz network in which the cascade depth grows rapidly indicating that the information diffuses *slowly* until the opinion threshold reaches approximately 0.1, after which the cascade depth appears to decay almost exponentially. We expect that this behaviour is attributable to the high-clustering of the network.

The small-world property means that a nodes "friends" are likely to also be friends, creating clusters of interconnected nodes within the network. As a result when the opinion threshold is small many neighbouring nodes are able to continue the cascade in successive iterations, whilst when the opinion threshold increases the effects of clustering become more prominent and allow for many nodes to be added to the cascade in one step: this reduces the number of iterations.

Amongst the remaining topologies we can broadly see two patterns emerging. We see that the SNAP Facebook and SNAP Twitter networks gradually increase in cascade depth with increasing threshold whilst the directed Barabási-Albert, standard Barabási-Albert and Erdős-Rényi networks tend to decay in depth. For the SNAP networks this behaviour appears to reinforce the expected behaviour for a real-world network: as you increase the openness for nodes to communicate the cascade depth which can also be thought of as the cascade *lifetime* increases. This is explained by the properties of these example networks as was discussed in chapter 3. Whilst both networks exhibit some clustering it is not as strong as that observed in the artificially generated Watts-Strogatz network. Extremely high clustering leads to an abundance of hub nodes in the network meaning that information is able to diffuse faster across the whole network because it is easier to reach a hub node. The lower clustering exhibited in the SNAP networks explains why the cascade continues further with increasing threshold, because there are fewer hub nodes and so further steps are needed. The greater depth for the SNAP Facebook network can be explained by its larger size.

The similar trend between the cascade depths of the directed Barabási-Albert, standard Barabási-Albert and Erdős-Rényi networks can also be analysed with similar reasoning. As examples of scale-free networks the Barabási-Albert graphs contain a small number of nodes with very high degrees with the remaining nodes having quite small degrees. This property as generated using the preferential attachment mechanism means that the majority of "smaller" nodes will follow at least one of the highly *influential* nodes. Because of this it takes very few iterations for information to diffuse: this is particularly clear for the standard undirected Barabási-Albert network.

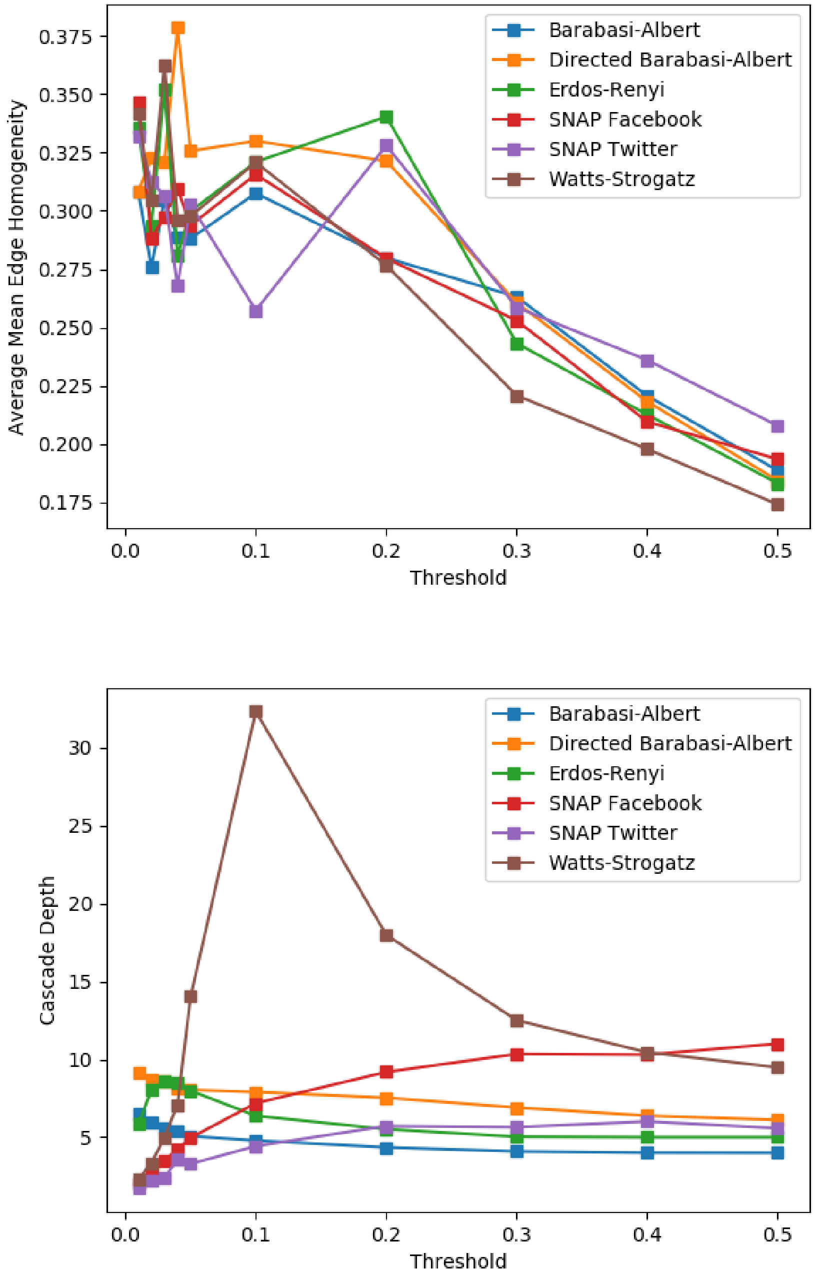


Figure 6.2: Comparison of the average mean edge homogeneity and average cascade depth as functions of the opinion threshold across different network topologies.

6.3 Role of Opinion Distribution

6.3.1 Motivation and Setup

We will experiment with 3 different opinion distributions across the range of different network topologies discussed above. We now motivate the inclusion of each of these distributions.

- The *uniform distribution*, more specifically a $U(0, 1)$ distribution, was used during our initial model behaviour evaluation and we will continue to use it as a baseline when comparing to other network topologies and distributions.
- We distribute opinions *normally* according to an $N(0.5, 0.25)$ distribution. We determined the μ and σ^2 parameters through inspection to obtain a suitable standard bell-shaped curve in the interval $[0, 1]$. We include this distribution to test our model mechanics under the assumption that most users in a social network have an "average" opinion, despite the observations of polarisation in the literature.
- Finally we distribute the node opinions according to a *bimodal* distribution to simulate examples of networks in which individuals are extremely polarised. We achieve this by sampling a $B(0.5, 0.5)$ distribution.

For the purposes of this experiment we will focus our discussion on the results for the SNAP Facebook and SNAP Twitter topologies: together these example networks provide the best representation of real social networks and provide the basis for a good comparison between a highly-clustered undirected network and a less-clustered directed network. We also focus on analysing cascade results for small opinion thresholds to ensure we best capture the stochastic element of the model.

6.3.2 Analysis

Due to the way in which edge homogeneity is defined and the properties of the different distributions we are using to model user opinion we can already deduce some expected behaviour before even looking at our simulation results. All three of the distributions being investigated have an expected value of 0.5 and so given our definition of mean edge homogeneity described in section 4.4.2 we expect a peak in probability density around 0, similar to that observed during our initial model evaluation in section 4.4. We would expect this to be most prominent for the $N(0.5, 0.25)$ distribution with low probability density for higher homogeneities. The bimodal property of the $B(0.5, 0.5)$ distribution should result in a higher peak of probability density close to 1.

The average mean edge homogeneity as a function of the opinion threshold is of little interest in this experiment as the relative value of the mean edge homogeneity is largely determined by the behaviour of the opinion distributions. As a result it can be qualitatively described without observing any experimental results: the quantitative results provided by experimentation gives little extra insight. Across all opinion thresholds the bimodal distribution gives the largest average mean edge homogeneity, followed by the uniform and then normal distribution. Additionally all three distributions follow the same slight downward trend in average mean edge homogeneity for increasing threshold, which is expected as a larger proportion of all edges in the underlying network are included. We also find that the cascade depth provides little insight as all three opinion distributions follow largely the same trend that is identical to that shown in figure 6.2.

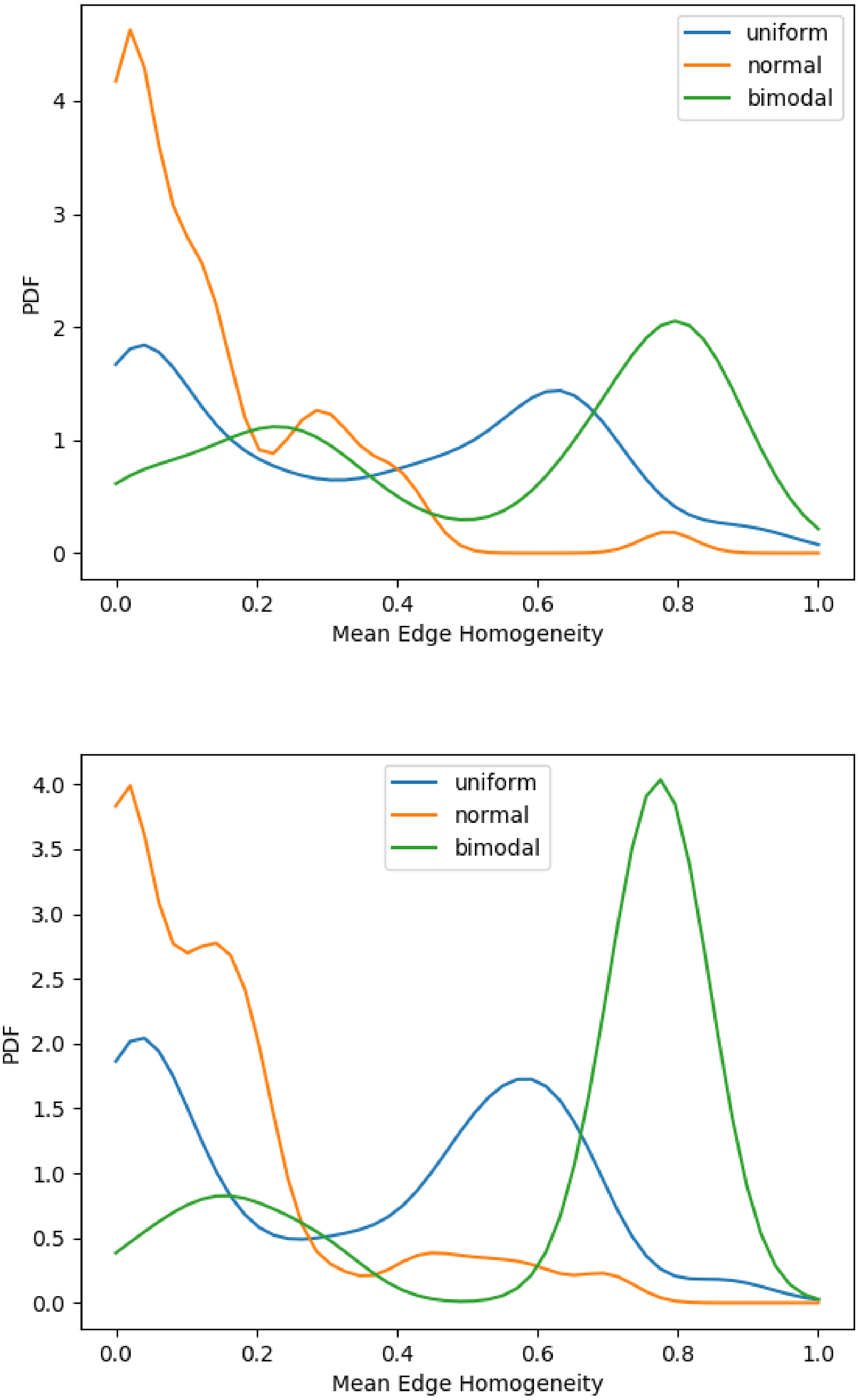


Figure 6.3: Comparison of the PDF of mean edge homogeneity across different opinion distributions for both the SNAP Facebook (top) and SNAP Twitter (bottom) example networks. These results were obtained from cascade simulations with the opinion threshold set to 0.01.

Figure 6.3 shows the PDF for mean edge homogeneity across the three different opinion distributions for both the SNAP Facebook and SNAP Twitter networks with opinion threshold set to 0.01. We can see that for both differing network topologies the PDF structure broadly follows the expected trends described qualitatively, with no observed instances of negative mean edge homogeneity in the simulations. This again shows that the majority of neighbouring nodes taking part in the cascade are homogeneous.

We see a steady decay in probability density when the opinions of nodes in the underlying network are normally distributed: this is intuitive as most node opinions are close to 0.5. For uniformly distributed opinions it is interesting to observe that there are nearly identical peaks around 0.05 and 0.6. The sharp peak around 0.8 for bimodally distributed opinions also confirms our predictions prior to experimentation, however it is interesting to observe that the peak is significantly sharper for the SNAP Twitter network. We hypothesise that this may be due to the directed nature of the network: if many nodes "follow" each other then this essentially results in each homogeneous edge being counted twice, pushing the probability density up. Further analysis of the underlying SNAP Twitter network could confirm this.

6.4 Role of Seed Node in Scale-Free Networks

6.4.1 Motivation and Setup

One key factor that we are yet to consider is the choice of seed node, which can clearly have some effect on cascade dynamics. The effects of this choice depends heavily on the properties of the underlying social network. Some topologies such as scale-free or so-called "influencer" networks have a clearer distinction between types of nodes that may affect the cascade: the scale-free property describes how a small number of nodes are highly connected whilst the majority of nodes have low degrees. This partition of the set of nodes into two sets with vastly different degrees creates the opportunity for the choice of seed node to potentially have a large effect on the cascade dynamics.

For this reason we limit our investigation into the role of this factor to scale-free networks. An argument could be made that the notion of "hub" nodes within small-world networks also provides opportunity for experimentation however the high-clustering of these networks (particularly in our examples) means that it is highly likely that most nodes are connected to a "hub" node anyway.

For this part of the investigation we will use two different distributions for the selection of a seed node. We now describe and motivate the inclusion of each of them.

- We select a seed node from the set of all nodes *uniformly* i.e. where each node has an equal probability of being selected.
- We also select a seed node according to the degree distribution of the network, thereby weighting the probabilities so that a higher degree node in the network is more likely to be chosen as the seed node than a lower degree node. This allows us to simulate the situation in which information more often originates from a popular or *influential* individual.

6.4.2 Analysis

As introduced above we only consider the scale-free Barabási-Albert network topologies (both directed and undirected), and we focus our attention to the case where opinion is distributed uniformly.

Unsurprisingly the PDF of mean edge homogeneity is hardly distinguishable between the two distributions for selecting a seed node particularly as the opinion threshold increases above 0.2: a similar behaviour applies to the average mean edge homogeneity as a function of opinion threshold. For this reason we focus our analysis on the depth of the cascade which offers more interesting insights.

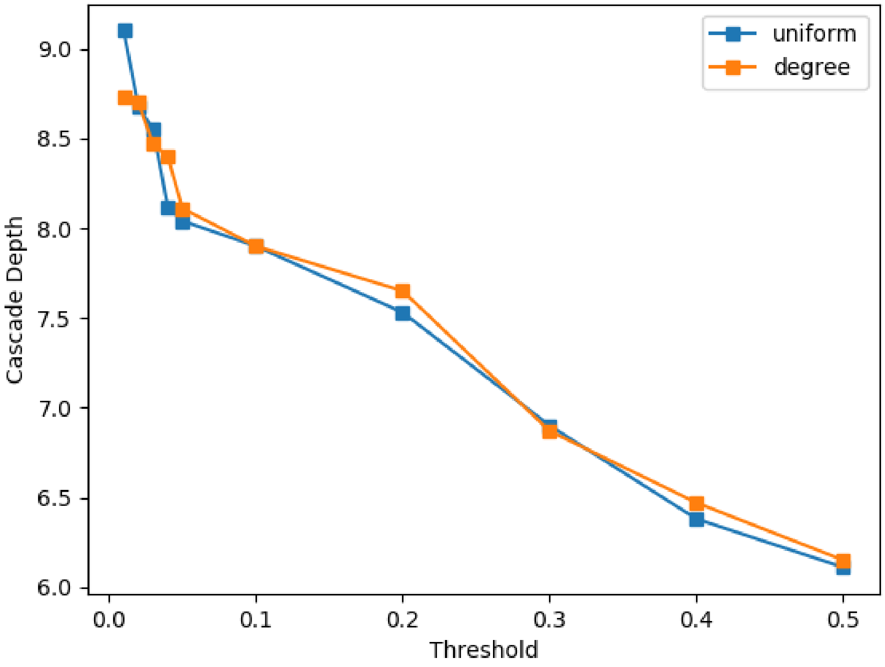
Figure 6.4 shows the average cascade depth as a function of the opinion threshold for both scale-free network topologies when the seed node is chosen using a uniform and degree-based distribution. Whilst the two plots representing the different topologies clearly exhibit different behaviours it is interesting to observe that the choice of seed node appears to have little impact on the cascade dynamics.

We hypothesise that this is due to the structure of scale-free networks. The preferential attachment mechanism used in the generation of Barabási-Albert networks leads to the creation of networks with low clustering and low average path length, with a small number of nodes having large degrees and the majority of nodes having low degree. This means that even if the chosen seed node is not an *influencer* node i.e. one with high degree that will diffuse information quickly, it is very likely that the seed node will be connected to such a node. Therefore within a small number of steps the cascade can reach an *influencer* node and spread rapidly: this is even more true for the undirected network as can be seen in the top graph of figure 6.4. This behaviour might also explain why we see negligible difference between the two different seed node distributions as the number of nodes with a direct link to a high degree node is so large that it makes little difference whether the cascade actually originates at a high degree node.

6.5 Evaluation

In section 4.4 we evaluated our model behaviour in relation to the core findings from the literature that information spreads more rapidly between like-minded individuals. In this section we look to briefly evaluate our experimental findings in comparison to our goals to consider what effects other factors can have in information diffusion. Whilst this is quite an open goal, we hope to be able to relate our findings to some conclusions from the literature to further validate them.

Throughout this chapter we have considered the role that network topology and the choice of seed node can have on the information cascade, in comparison to the role of node opinions which we are using to model social products. We have shown through simulations using our model that by far the factor that has the biggest influence on cascade dynamics is the opinion of nodes: this behaviour closely resembles findings from the literature. Vicario et al. showed in work such as [6] that highly polarising information such as conspiracy theories spreads much faster than less controversial claims such as scientific news. By analysing the diffusion of conspiracy theories on Facebook Vicario et al. showed that large conspiracy cascades often elicited high levels of mean edge homogeneity in the



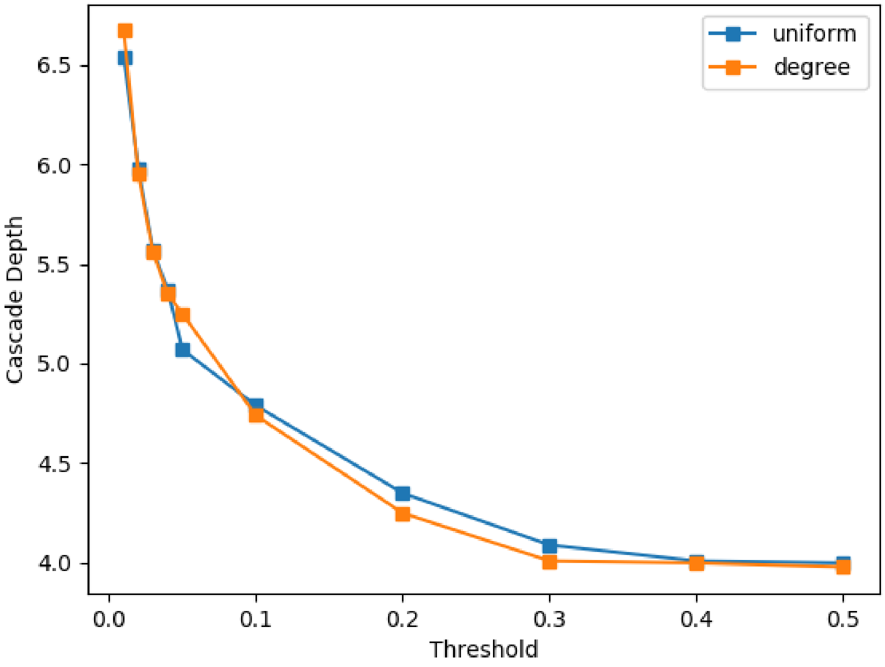


Figure 6.4: Comparison of the average cascade depth between a uniformly and degree distributed choice of seed node for both the directed Barabási-Albert network (top) and standard Barabási-Albert network (bottom).

interval $[0.5, 0.8]$. We can observe similar behaviour with our model when node opinion is distributed bimodally to simulate a network with highly polarised individuals.

6.5.1 Strengths

- The agreement between our findings that social products play the largest role in cascade dynamics and conclusions made in the literature verifies that our model can represent real-world cascades quite well.
- The configurability of our model allows it model a wide range of possible scenarios.
- Our experiments also show that network topology does play a role in information cascade dynamics however its effects are heavily tied to the opinions of the connected nodes.

6.5.2 Weaknesses

- We have not been exhaustive in our experimentation: there are potentially other factors that we have not considered here that may have significant effects on cascade dynamics. We will discuss this further in section 7.2.3.
- Not all of our findings are verifiable against the literature. For example we are unable to compare our findings regarding network topology quantitatively to real analysis: we can only qualitatively come to similar conclusions.

Chapter 7

Conclusion and Future Work

7.1 Summary

Throughout this project our main objective of building a new information cascade model that focusses on the diffusion of information by *social* means has been met. Influenced and motivated by the literature we have used similar mechanics to model information diffusion from a new angle: where the literature has focussed on interaction between individuals and news sources we have focussed on peer-to-peer diffusion.

To test our model, we studied and collected different network topologies and found ways to generate example networks. As part of this work we adapted the pre-existing SNAP datasets to better fit our use cases. To aid our analysis and model evaluation we have built both static and interactive graph visualisation tools.

To evaluate our model we compared its behaviour both qualitatively and quantitatively against that observed in real-world studies and other information cascade models in the literature. We found that our model behaves in broadly the same way and displays outcomes that are compatible with those previously observed in the literature by comparing the depth and mean edge homogeneity of cascades. We have then used our model to investigate other contributing factors to information cascades online, and found that the distribution of node opinions has the biggest effect on cascade dynamics. These node opinions are intended to model a spectrum of social products that individuals could possess.

By designing, implementing and experimenting with this model we have approached the topic of information cascade modelling from a different angle to that previously studied. When combined with existing models we are now able to model a wide range of scenarios and mechanics for how individuals use social media. The existence of accurate models for online social media use allows predictions to be made about how information might disseminate in the future, and therefore how misinformation in particular might be controlled. Our experimentation has shown that the distribution of social products in the network plays a large role, and that in particular a bimodal distribution i.e. a network of strongly *polarised* individuals results in the largest and potentially most dangerous information cascades. These findings reinforce those found by Vicario et al. in [6] and highlights again that polarisation and controversy are important factors in information diffusion.

Our experimentation also shows that network structure and topology play a role in cascade dynamics, although to a lesser extent than opinion distribution. We have shown in particular that certain well-connected nodes in scale-free networks have a very large *influence* on the cascade. These findings highlight that identifying key *influential* individuals in a network could be an effective tactic in the fight against misinformation online, and links our work into wider studies such as that presented in [32].

7.2 Extensions

Whilst our project has largely met its intended goals and objectives there are multiple directions in which our work could be taken in future given more time and scope.

7.2.1 Expanded Model Mechanics

As we have emphasised throughout this report, our project has focussed on modelling the *social* transmission of ideas i.e. information being passed directly from one individual to another through online social networks. This has been our focus with the aim of augmenting much of the work in the literature, which has focussed on users consuming news and information from news outlets through the medium of online social media.

Clearly these ideas are related, and as we have discussed we have used similar mechanics to model both scenarios. Therefore in future it could be interesting to combine elements of both models to try and simulate the full spectrum of actions possible on online social media services. It is important to consider the goal this would aim to achieve as models and simulations tend to be simplistic for a reason. However a more complex model would potentially be applicable and configurable to a wider range of scenarios.

7.2.2 More Realistic Network Examples

Throughout this work we have had to make some quite strong assumptions about the way online social networks are structured and the opinions, beliefs and habits that individuals using these services have. This is part of the process of designing any mathematical model: scenarios are usually simplified and assumptions are made to allow progress to be made. Whilst the assumptions we have made have been sensible and based on analysis of real social networks, with more time and scope we think that we could further refine our model assumptions and hopefully our model behaviour in the process. For example with more time and resources we could potentially look to analyse real social media usage to capture user sentiment, with the aim of gaining a deeper understanding of how user opinion is distributed within online social networks. Such work would require substantially more resources and time, and would of course need to be conducted ethically and legally however it could be a valuable and interesting piece of future work to better understand the sociology of user's behaviour online and allow us to improve our model assumptions in the process.

We did briefly consider going down this path as part of our work, as on initial inspection it appeared that the SNAP datasets detailed in chapter 3 also came with anonymised node *features*. Initially we thought that we might be able to map these node features to node opinions for our model, however the dataset was incredibly sparse and incomplete: not every node had features, and not every node had the same features, making it incredibly difficult to standardise.

7.2.3 Further Experimentation

In chapter 6 we focussed on the key factors that could affect information cascade simulations using our model (and in the real-world): user opinions, network topology and we briefly considered the choice of seed node. Whilst we feel that these are clearly the largest factors and those most likely to meaningfully affect cascade dynamics, we have by no means been exhaustive in our experimentation. Perhaps the largest omission from our investigation is the wider consideration of the role of seed nodes. Possible pathways for experimentation include varying the number of seed nodes and the way in which they are chosen. Similar experimentation was used during the evaluation of the cascade model proposed in [6] and so it could be interesting to see the effects that this factor would have on our model behaviour.

7.2.4 More Integrated Pipeline

Our work in this project has focussed heavily on the model design, experimentation and evaluation. Whilst we have created some visualisation tools to aid in our analysis and model development these are quite limited in scope. Given more time it would be beneficial to improve these visualisation tools and better integrate them with the model creation and simulation steps of our program. We envisage a single application or dashboard that allows users to create models, run simulations and visualise them simultaneously and interactively within one window. This would allow for more immediate feedback on experimentation results and simply provide a tidier interface to our model.

Bibliography

[1] Philip Ball. The epidemiology of misinformation, 2020. URL https://www.prospectmagazine.co.uk/science-and-technology/epidemiology-misinformation-coronavirus-covid19-conspiracy-theory. [Accessed: 20th May 2020].

[2] Lee Howell. Digital wildfires in a hyperconnected world. Technical report, World Economic Forum, 2013. URL http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/?doing_wp_cron=1579170965.0308299064636230468750. [Accessed: 21st January 2020].

[3] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. 2020. URL https://arxiv.org/abs/2003.05004. [Accessed: 18th May 2020].

[4] Jigsaw Research. News consumption in the uk: 2019 report, 2019. URL https://www.ofcom.org.uk/__data/assets/pdf_file/0027/157914/uk-news-consumption-2019-report.pdf. [Accessed: 18th May 2020].

[5] Walter Quattrociocchi, Antonio Scala, and Cass R. Sunstein. Echo chambers on facebook. June 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795110. [Accessed: 22nd January 2020].

[6] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, 113(3):554, Jan 19, 2016. URL https://www.pnas.org/content/113/3/554. [Accessed: 21st January 2020].

[7] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. ISSN 1476-4687. doi: 10.1038/30918. URL https://doi.org/10.1038/30918. [Accessed: 14th April 2020].

[8] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. URL https://books.google.co.uk/books?hl=en&lr=&id=CAm2DpIqRUIC&oi=fnd&pg=PR21&dq=Social+Network+Analysis:+Methods+and+Applications.&ots=HwOnscWATd&sig=wLgIO7UlMG5csHpCFnrz-i4NFIw&redir_esc=y#v=onepage&q=Social%20Network%20Analysis%3A%20Methods%20and%20Applications.&f=false. [Accessed: 13th June 2020].

[9] J. Bouttier, P. Di Francesco, and E. Guitter. Geodesic distance in planar graphs. *Nuclear Physics B*, 663(3):535–567, Jul 2003. ISSN 0550-3213. doi: 10.

1016/s0550-3213(03)00355-9. URL http://dx.doi.org/10.1016/S0550-3213(03)00355-9. [Accessed: 13th June 2020].

[10] Oxford University Press. Lexico.com. URL https://www.lexico.com. [Accessed: 11th June 2020].

[11] Roy M. Anderson and Robert M. May. Population biology of infectious diseases: Part i. *Nature*, 280(5721):361–367, 1979. ISSN 1476-4687. doi: 10.1038/280361a0. URL https://doi.org/10.1038/280361a0. [Accessed: 24th January 2020].

[12] Paulo Shakarian, Abhinav Bhatnagar, Ashkan Aleali, Elham Shaabani, and Ruocheng Guo. *Diffusion in Social Networks*. Springer International, 2015. ISBN 978-3-319-23104-4. URL https://link.springer.com/chapter/10.1007/978-3-319-23105-1_4. [Accessed: 24th January 2020].

[13] Paul Erdos and Alfred Renyi. *Publicationes Mathematicae Debrecen*, pages 290–297. URL https://www.renyi.hu/~p_erdos/1959-11.pdf. [Accessed: 27th April 2020].

[14] Jon Kleinberg. Complex networks and decentralized search algorithms. URL https://www.stat.berkeley.edu/~aldous/Networks/icm06.pdf. [Accessed: 27th April 2020].

[15] Jason Wohlgemuth and Mihaela Matache. Small-world properties of facebook group networks. *Complex Systems*, 23:197–225, 09 2014. doi: 10.25088/ComplexSystems.23.3.197. URL https://www.researchgate.net/publication/289259407_Small-World_Properties_of_Facebook_Group_Networks. [Accessed: 27th April 2020].

[16] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.509. URL https://science.sciencemag.org/content/286/5439/509. [Accessed: 16th April 2020].

[17] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002. doi: 10.1103/RevModPhys.74.47. URL https://link.aps.org/doi/10.1103/RevModPhys.74.47. [Accessed: 28th April 2020].

[18] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. Digital news report. Technical report, Reuters, 2019. URL http://www.digitalnewsreport.org. [Accessed: 27th January 2020].

[19] Michela Del Vicario, Fabiana Zollo, Guido Caldarelli, Antonio Scala, and Walter Quattrociocchi. Mapping social dynamics on facebook: The brexit debate. *Social Networks*, 50:6–16, Jul 2017. doi: 10.1016/j.socnet.2017.02.002. URL https://www.sciencedirect.com/science/article/pii/S0378873316304166. [Accessed: 21st January 2020].

[20] Aris Anagnostopoulos, Alessandro Bessi, Guido Caldarelli, Fabio Petroni, Michela Del Vicario, Antonio Scala, Fabiana Zollo, and Walter Quattrociocchi. Viral misinformation: The role of homophily and polarisation. 2014. URL https://arxiv.org/pdf/1411.2893.pdf. [Accessed: 27th January 2020].

[21] Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. Anatomy of news consumption on facebook. *Proceedings of the National Academy of Sciences*, 114(12):3035–3039, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1617052114. URL https://www.pnas.org/content/114/12/3035. [Accessed: 2nd February 2020].

[22] Guillaume Deffuant, David Neau, Frederic Amblard, and Gerard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000. ISSN 1793-6802. URL https://www.worldscientific.com/doi/abs/10.1142/s0219525900000078. [Accessed: 2nd February 2020].

[23] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, jun 2014. [Accessed: 28th April 2020].

[24] Jure Leskovec and Julian J. Mcauley. Learning to discover social circles in ego networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 539–547. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4532-learning-to-discover-social-circles-in-ego-networks.pdf. [Accessed: 28th April 2020].

[25] Geoff Sims. Function to produce bezier curves for the edges in a networkx graph, 2019. URL https://github.com/beyondbeneath/bezier-curved-edges-networkx. [Accessed: 29th April 2020].

[26] Networkx: Network analysis in python. URL https://networkx.github.io. [Accessed: 21st May 2020].

[27] Pandas python library. URL https://pandas.pydata.org. [Accessed: 21st May 2020].

[28] Matplotlib: Visualisation with python. URL https://matplotlib.org. [Accessed: 21st May 2020].

[29] Dash python library. URL https://dash.plotly.com. [Accessed: 21st May 2020].

[30] J H Wang. Network-visualisation. URL https://github.com/jhwang1992/network-visualization. [Accessed: 16th May 2020].

[31] Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, page 132–139, USA, 2003. Society for Industrial and Applied Mathematics. ISBN 0898715385. URL https://dl.acm.org/doi/10.5555/644108.644133. [Accessed: 29th May 2020].

[32] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012. ISSN 0036-8075. doi: 10.1126/science.1215842. URL https://science.sciencemag.org/content/337/6092/337. [Accessed: 22nd November 2019].

[33] Product distribution of two uniform distributions. URL https://math.stackexchange.com/questions/659254/product-distribution-of-two-uniform-distribution-what-about-3-or-more. [Accessed: 24th April 2020].

Appendices

Appendix A

Discussion of Threshold Effect on Mean Edge Homogeneity

In section 4.4.3 we mention how our initial simulations in which the node opinions are uniformly distributed show that as the opinion threshold goes to 1, the mean edge homogeneity goes to 0. Figure A.1 shows this behaviour as observed in our simulations.

A.1 Intuition

In these simulations the node opinion in the underlying network is distributed according to a $U(0,1)$ distribution. This means that the polarisation σ_u of node u defined in section 4.4.2 is distributed according to $U(-1,1)$. Therefore the expected value of polarisation is 0, and so across the whole underlying network the mean edge homogeneity will be 0. The threshold condition is designed to encourage similar nodes with closer polarisations to interact more than less similar neighbours, however as the threshold increases this threshold becomes redundant as all neighbouring nodes in the underlying graph are included in the cascade. This causes the mean edge homogeneity to go to 0 in the cascade.

A.2 Proof

Let X_u and X_v be two independent and identically distributed random variables representing the opinions of two neighbouring nodes u and v in a network. In these simulations we have $X_u \sim U(0,1)$ and $X_v \sim U(0,1)$. Let Z be a random variable denoting the edge homogeneity distribution. From section 4.4.2 we know that the edge homogeneity is defined as the product of the node polarisations. This is given by:

$$Z = X_u \times X_v = (2X_u - 1) \times (2X_v - 1) = 4X_uX_v - 2X_u - 2X_v + 1 \tag{A.1}$$

and we want to determine $E(Z)$. We can now apply the linearity property of expectation.

$$\begin{aligned} E(Z) &= E(4X_uX_v - 2X_u - 2X_v + 1) \\ &= 4E(X_uX_v) - 2E(X_u) - 2E(X_v) + 1 \end{aligned}$$

By the properties of the uniform distribution, we already know that $E(X_u)$ and $E(X_v)$ are equal to $\frac{1}{2}$. Therefore we require $E(X_uX_v)$. Let us define $Z_1 = X_uX_v$. The cumulative

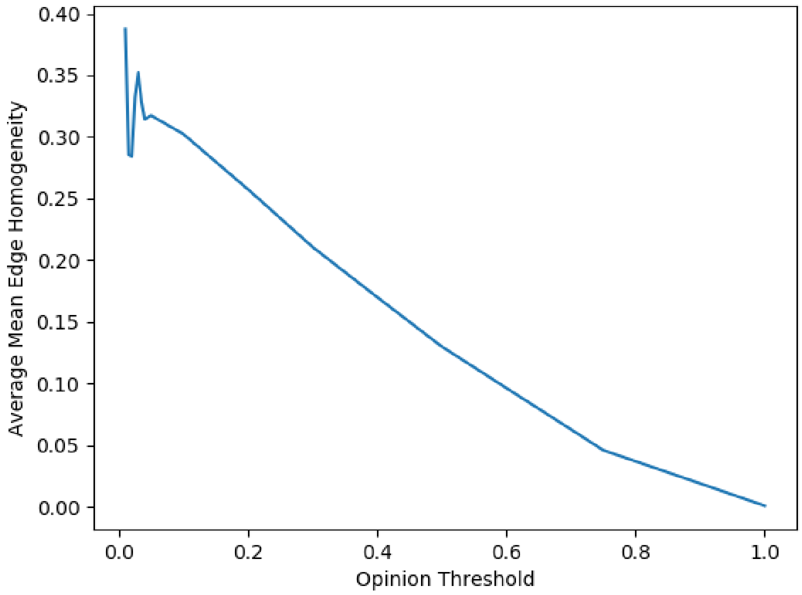


Figure A.1: The average mean edge homogeneity of 100 cascade simulations as a function of the parameterised opinion threshold.

distribution function of Z_1 is then derived as follows as first shown in [33].

$$\begin{aligned} F_{Z_1}(z) = P(Z_1 \leq z) &= \int_{x=0}^{1} P(X_v \leq \frac{z}{x}) f_{X_u}(x) dx \\ &= \int_{x=0}^{z} dx + \int_{x=z}^{1} \frac{z}{x} dx \\ &= [x]_0^z + z\,[\ln(z)]_z^1 \\ &= z - z\ln(z) \end{aligned}$$

Therefore the density function of Z_1 is $f_{Z_1}(z) = -\ln(z)$. We can now determine the expectation of Z_1.

$$\begin{aligned} E(Z_1) &= \int_0^1 -z\ln(z)dz \\ &= -\int_0^1 z\ln(z)dz \\ &= -\left[\frac{1}{2}z^2\ln(z) - \frac{z^2}{4}\right]_0^1 \\ &= \frac{1}{4} \end{aligned}$$

Plugging this expectation back into our original equation for the expectation of Z gives us:

$$E(Z) = 4\frac{1}{4} - 2\frac{1}{2} - 2\frac{1}{2} + 1 = 0 \tag{A.2}$$

which matches our experimental observations.

Appendix B

Directed Barabási-Albert Network Degree Distributions

In section 6.2.1 we described a new process that uses the existing Barabási-Albert model for scale-free graph generation to generate *directed* graphs that are approximately scale-free. Below we show the in-degree and out-degree distributions for an example model generated using this process. These distributions clearly show an approximate scale-free structure in which the in-degree and out-degree distributions follow different power laws (i.e. power laws with different exponents).

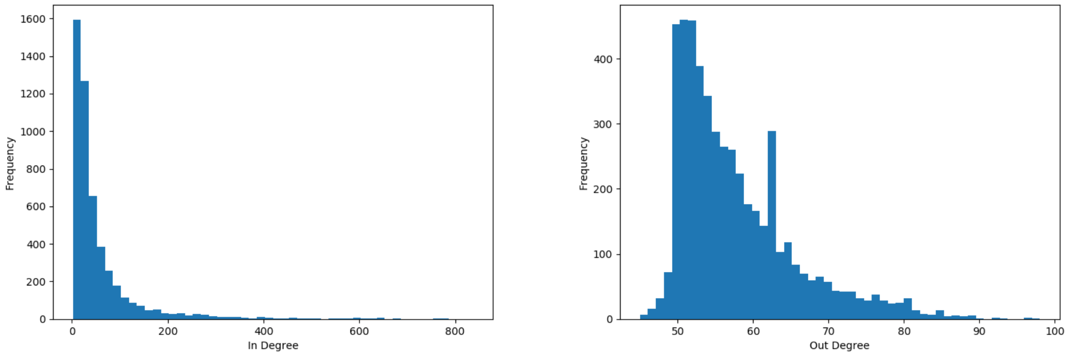


Figure B.1: Histograms showing the in and out degree distribution for an example directed graph generated using the process described in section 6.2.1. The underlying Barabási-Albert model used parameters $n = 5000, m = 50$.