Using Agent-Based Modelling to Investigate Intervention Algorithms to Reduce Polarisation in Online Social Networks

A. M. Coates

PhD 2020

Using Agent-Based Modelling to Investigate Intervention Algorithms to Reduce Polarisation in Online Social Networks

Manchester Metropolitan University

Faculty of Science and Engineering
Department of Computing and Mathematics

Thesis submitted in partial fulfilment of the
of the requirements of the degree of

Doctor of Philosophy

Adam Coates

2020

Abstract

Across much of the western world, political polarisation is on the rise. This has the effect of hindering political discourse, stifling open discussion, and in extreme cases has led to violence. The process of polarising and radicalising vulnerable individuals has migrated to social media websites, which have been implicated in several high profile terror attacks.

Within this thesis we model and investigate various algorithms to prevent the spread of polarisation and extremist ideology by employing agent-based modelling techniques from the field of opinion dynamics. The contributions of our work include the following aspects.

Firstly, we have developed a unified framework for opinion dynamics, allowing us to experiment easily on a number of different existing models and bringing together sometimes disparate innovations from across the field into one system.

Secondly, this unified framework has been implemented in a modular simulator able to perfectly replicate results from purpose-built, stand-alone simulators for two widely used models, namely Relative Agreement and CODA, and then released to the public as the first general-purpose opinion dynamics simulator.

Thirdly, we have developed two new intervention algorithms, along with a new metric for measuring the effectiveness of an intervention strategy, which aim to reduce the spread of polarisation across a network with low computational cost. These methods are compared to existing centrality-based methods upon a random network. The experimental results show our proposed approaches outperform centrality measures. We find that our

algorithms are able to prevent up to 40% of non-extremist agents becoming extreme by removing only 10% of the network's edges.

Fourthly, we have investigated the efficacy of these intervention algorithms on polarisation under different scenarios (e.g. variable costs, different network structures). The experimental validation proves the proposed approach is robust and has performed favourably compared existing methods such as centrality-based methods especially on the second type of network.

Finally, we have developed a broadcast-based communication system for agents, designed to mimic the one-way broadcast nature of a public social media post such as Twitter, in contrast to the existing model which emulates a two-way private conversation. The experimental result shows a lessening of the impact of our interventions, demonstrating the need for further investigation of such communication methods.

Acknowledgements

First and foremost, my sincere gratitude to Dr Anthony Kleerekoper, for his continuous guidance and support. His knowledge of and passion for the subject was a source of inspiration throughout my studies, and I couldn't have asked for a better mentor. His ideas, patience, and constant assistance made this research possible.

I would also like to thank Prof Liangxiu Han, Dr Emma Norling, and Prof Jon Bannister for their assistance, advice, and comments during my research. Without their support, this project would similarly have been impossible.

My thanks and appreciation also go out to my fellow research students for their discussions, the IT support staff of the university for their timely technical assistance in times of need, and the research degrees administration staff for their quick and helpful response to any queries I had.

I'd finally like to thank my family, for their unconditional love and support, and providing a place to stay when I needed a break from the stress. And last but certainly not least, my wonderful Olivia, whose presence by my side helped more than she will ever admit.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and Background

In recent years, political polarisation and radicalisation has led to a breakdown in political discussion (Hong and Quarterly, 2016; Mendez, Cosby, and Mohanty, 2018; Richter, 2019) and an increase in domestic terrorism (Quek, 2019). An analysis of 679,000 Twitter users over 8 years showed that political polarisation has increased between 10% and 20% (Garimella, 2017), while a separate analysis of over 1,000,000 users showed that users voluntarily segregate themselves into distinct clusters around a political issue such as Brexit (Vicario, Zollo, and Caldarelli, 2017). In such segregated groups users are only exposed to information they are likely to agree with, which then reinforces and strengthens their beliefs. Additionally, governmental figures are increasingly becoming more polarised and extreme in their professed beliefs, which in turn increases their follower count, reach, and ability to radicalise others (Hong and Quarterly, 2016; Baldassarri and Gelman, 2008).

2019 saw an unprecedented combination of terrorism and social media in the form of a deadly shooting spree in Christchurch, New Zealand. On the 15th of March, 2019, the attacker first posted a manifesto to the anonymous image-sharing website 8chan and

emailed it to several government figures and media outlets, and then began shooting people at a mosque and an Islamic centre while streaming the footage over Facebook Live. Since then, several similar attacks have been carried out using similar methods, ideology, and apparent method of radicalisation such as in El Paso and Poway. Quek noted this and stated "the El Paso attack is the third mass shooting in 2019 within the US, linked to the online forum 8chan and one of several recent attacks committed by individuals to credit the Christchurch attack as an inspiration." (Quek, 2019). Quek also noted that unlike many terror attacks however, in each of these cases the perpetrators have had no prior criminal history or outstanding mental health issues that might otherwise have alerted law enforcement or intelligence services to their intentions. In addition to taking place in meeting houses, places of worship, or prisons, political polarisation is now increasingly taking place entirely online, making methods to identify and combat it all the more important (Fernandez, Asif, and Alani, 2018).

Combating online polarisation and radicalisation has been tried with many approaches and met with mixed success (Home Office, 2018; Fernandez, Asif, and Alani, 2018; Wright, Graham, and Jackson, 2017). While the administrators of some social media sites make an attempt to censor extremist material and ban extremist accounts, Facebook and Twitter have both stated that their policies against extremist material are not applied to world leaders (BBC News, 2019b; Twitter, 2019). Further to this, several social media sites and discussion boards are well-known for their reluctance to perform moderation, and have seen frequent use by extremists. This includes the far-right social networking site Gab, which bills itself as "a social network that champions free speech, individual liberty and the free flow of information online" (Webster, 2019), as well as the so-called "Politically Incorrect" sections of anonymous image-sharing websites 4chan and 8chan (Hine et al., 2017; Evans, 2019)[1]. Due to the anonymity and lax moderation of 4chan and 8chan, they have become refuges for fringe political groups, hacktivist efforts, and

[1]These three sites should be considered extremely not-safe-for-work

speech considered unacceptable elsewhere.

Third parties have attempted to perform some intervention on the spread of polarised ideology, such as the UK government's Prevent strategy's approach of identifying and intervening directly with vulnerable individuals (Home Office, 2018), and the acts of a movement of computer hackers known as Anonymous, who acted after the Paris attacks of November 2015 to hack over 20,000 Twitter accounts, remove Islamic State propaganda from the accounts, and frequently upload homosexual pornography in its place (Fernandez, Asif, and Alani, 2018). However, given the growth in polarisation despite all of these methods, it is clear a new solution is required. Opinion dynamics research in this area is limited, as shown in section 2.4. We also note that the proposed solutions are of high computational complexity, and therefore likely impractical for use on social networks with a very large number of users.

In this thesis, we investigate methods by which networks can be protected against the influence of extremists. As it is typically easy to create a new account should one be banned, we investigate ways in which polarisation can be prevented without resorting to banning. Instead, we look at ways that communications can be limited between radical and vulnerable people or accounts in such a way as to reduce the spread of polarising influence.

To model such networks and the competition of mutually exclusive influences, we turn to opinion dynamics. This is a form of modelling in which a network of agents with simple attitudes and behaviours is created, and then these agents are left to interact with one another according to these behaviours until some form of stability is observed. As the kind of data required by this form of modelling is difficult or impossible to gather in real life such as objective numeric values of an actor's opinion, we exclusively use simulated data. Within opinion dynamics this is however not a limitation: models are expressly designed to abstract away much of the complexity that comes with real-world data, allowing us to focus exclusively on the fundamentals of a complex situation that

would be impossible to model exactly. Simulations grant us a level of scale and flexibility of approaches that would be impossible with real participants, yet can be used to inform later developments.

This field of agent-based modelling was originally devised to model the expansion and competition of species over a region of space (Clifford and Sudbury, 1973), but has since been expanded to cover the contagious behaviours of biological infections (Robinson, Cohen, and Colijn, 2012), rumours (Kimura, Saito, and Motoda, 2009), computer viruses (C. Gao et al., 2013), and extremism (Deffuant, Amblard, et al., 2002). There are now a multitude of different opinion dynamics models, which we review in chapter 2.

1.2 Research Questions

After a review of the literature, we identified four key research questions which we aim to answer within this thesis. These research questions are as follows:

- Research Question 1: Can the different opinion dynamics models be unified into a consistent framework? There are many existing opinion dynamics models, each focusing on a different aspect of how people come to form and hold opinions. Are there underlying features common to them that would highlight key assumptions that different models authors make, or are the models completely separate?

- Research Question 2: Can we implement this framework in a rigorous program? A theoretical framework helps us think about different models, but going one step further and producing a robust implementation allows us to conduct experiments in a methodical manner.

- Research Question 3: Can we design intervention algorithms that reduce the spread of polarisation? It is clear both that radicalisation through social media is an

increasing problem, and that existing attempts to prevent this are ineffective. We hold that this new approach may be more effective, and that it requires investigation.

- Research Question 4: Can we make simulations used in opinion dynamics more reflective of online social networks? To ensure that our methods are resilient to changes in network structure and communication method, we wish to make our simulation more realistic and ensure that our algorithms still remain effective.

1.3 Aim, Objectives, and Contributions

The overarching aim of this thesis is to explore options for interventions on social networks that can reduce the propagation of extremist opinion and thus polarisation, with limited changes to the network structure. We divide this aim into four objective.

- Objective 1: To develop a modular framework of independent components that can implement existing opinion dynamics models. Aside from being a prerequisite for our second objective, this framework will aid and enable collaboration between researchers in the future in both extremism research and further afield.

- Objective 2: To develop and release a simulator that implements this unified framework. This program will us to rapidly prototype and test algorithms, and to alter test conditions to ensure our algorithms are applicable to a wide variety of situations.

- Objective 3: To design new intervention algorithms for removing edges between agents and evaluate their effect on polarisation. This objective directly addresses the core of our aim, and will lead to the development of both our intervention algorithms and a thorough metric by which they can be tested.

- Objective 4: To investigate the applicability and efficacy of our intervention algorithms under different more realistic scenarios including variable edge costs, updated network structures and broadcast-based communication. This paves the way for future experimentation into social media-inspired network interventions.

1.4 Contributions

Our contributions are five-fold:

- Contribution 1: The first framework to unify models, streamline research for opinion dynamics across the field, expose new parameter space, and reduce duplication, enabling collaboration between researchers in the future in both extremism research and further afield (chapter 3).

- Contribution 2: The first general purpose agent-based modelling simulator targeted towards opinion dynamics (chapter 3), capable of perfectly replicating results from purpose-built, stand-alone simulators for two widely used models, namely Relative Agreement and CODA.

- Contribution 3: Two new intervention algorithms with low computational cost that substantially reduce polarisation within a network have been proposed, along with a new metric by which the efficacy of interventions can be judged (chapter 4).

- Contribution 4: Investigations of the efficacy of these two intervention algorithms under different scenarios by taking into account variable costs and different network structures (chapter 6).

- Contribution 5: A broadcast-based communication system for agents, designed to mirror many social networks such as Twitter for more realistic scenarios (chapter 6).

1.5 Thesis Structure

In chapter 2 we provide a background of opinion dynamics and an exploration of various methods already proposed to limit polarisation on networks.

Chapter 3 proposes a novel framework for opinion dynamics models that allows us to assemble existing or new models from independent "modules". This approach means we can rapidly test our methods on existing and new structures, and under various different assumptions for how agents communicate and interact with one another. The framework is then used to replicate two existing experiments to test and validate it.

In chapter 4 we introduce two new intervention algorithms, which aim to reduce the spread of polarisation across a network. These methods are compared to existing centrality-based methods upon a random network, and perform favourably.

Subsequent chapters then explore these algorithms in further detail, with alterations aimed to more closely represent real-world circumstances.

In chapter 5, we explore the efficacy of our intervention algorithms by taking variable costs into account such as giving edges different costs according to either the degree of that edge's agents, or the number of alternate paths of length two between that edge's agents - a proxy for the strength of the relationship between those agents. The extensive experimental evaluation is presented.

Chapter 6 explores the efficacy of our intervention algorithms under different network structures (e.g. Scale free network), along with extensive experimental evaluation. In section 6.4, we then altered the method of communication from individual one-to-one conversations to public broadcasts to all neighbours. This more closely mirrors the public posting nature of social networks like Twitter, in that by default all of your followers see your message at once.

Finally, we conclude in chapter 7 with a final observation of our results and a discussion on their meaning and implications, and some ideas for future developments in sections

7.1 and 7.2.

Chapter 2

Literature Review

In this chapter, we first give an overview of the field of opinion dynamics, our chosen approach to modelling extremism, radicalisation, and polarisation. This begins with a brief history, then moves on to a survey of some of the most popular and influential models used within the field, as well as some empirical studies using real-world data on opinions and relationships between humans. The key aim of our work is on preventing the spread of radical ideology, but as there is little research in this precise area we look at studies using similar models to prevent the spread of a biological pathogen. This provides us with inspiration for our own intervention methods as well as a guide to common limitations and pitfalls. Finally, we conclude this chapter with a discussion of concepts related to preventing ideological spread and a look at some of the most closely related work within this section.

2.1 Opinion Dynamics

Opinion dynamics is the study of how groups of individuals change opinions based on communication with one another. It is a type of agent-based modelling: a system in which individual agents are given states, properties, and behaviours, then placed within a

virtual environment and left to interact with each other and that environment. These small-scale micro-behaviours then give rise to large-scale macro-behaviours.

Opinion dynamics has its roots in cellular automata such as the Game of Life, developed by Conway in 1970 (see Gardner, 1970). In this game, each square in an infinite two-dimensional grid is considered in one of two states - "alive" or "dead". Rules are then introduced that govern the evolution of the grid over time. In each discrete time step, dead cells with three or more adjacent living cells come to life while living cells with one or fewer living neighbours die. The grid is seeded by setting a number of cells to be alive at the beginning of the simulation, and then the repeated application of these rules gives rise to complex mechanics.

Soon afterwards, a similar model was proposed by Clifford *et al.* to study conflict between two mutually hostile species over territory (Clifford and Sudbury, 1973). In this model, cells in a 2D lattice are in one of two states, occupied by one species or the other. The evolution of the system occurs in a non-deterministic manner: in each discrete time step, a random cell is chosen and it adopts the state of a randomly-chosen neighbour. The authors predicted that over time the model would stabilise into roughly circular regions, and were investigating whether or not it was possible to predict the duration of a conflict through observation of the territories controlled by each "side". The authors also noted that despite the underlying statistical mechanics obeying the law of the increase of entropy, their rules for modelling living things gave rise to the spontaneous emergence of order (Clifford and Sudbury, 1973). That is to say, from an initially random distribution of states, the simulations would always trend towards forming homogeneous "clumps" of cells in the same state.

From these initial beginnings, a large number of models rapidly emerged. It would be impossible to list every model used in opinion dynamics, but the following sections provide an overview of particularly influential models both to this project, and to the field in general. We first outline discrete models where opinions are held as a selection from a

Symbol	Description
i, j	Agents
s	Opinion
u	Uncertainty
s_i	The opinion of agent i
N	All agents within the network
τ	Assimilation threshold: agents within this range of a given opinion will be considered to agree with that opinion
ϵ	Repulsion threshold: agents outside this range of a given opinion will be hostile towards that opinion

Table 2.1: Symbols commonly used within this thesis.

list, before moving on to continuous models, where opinions are held as a real number. We then touch upon similar work within epidemiology. While not directly related to opinion dynamics, the models used by epidemiologists known as the SI family of models closely resemble the discrete models used in opinion dynamics, especially with regards to rumour spreading (Hosseini and Azgomi, 2016).

2.1.1 Mathematical Notation

Many of the models described here use similar or identical concepts within their mathematical representations. To aid the reader, these have been standardised in this thesis, and briefly summarised in table 2.1.

2.1.2 Discrete Models

Discrete models build upon their cellular automata heritage, owing in part to real-world applications and part due to the source of their inspirations. Some early models borrowed the idea of ferromagnetism from physics, in which an electron can have either an up spin or a down spin. Additionally, these models were initially used to analyse voting trends, in which voters might well hold infinitely variable opinions, but must select from a limited pool: either for or against, or for a specific political party (Katarzyna Sznajd-Weron and

Sznajd, 2001).

The Voter model was developed in 1975 to model how neighbours might persuade one another with regards to an abstracted voting process (Holley and Liggett, 1975). It functions identically to the model proposed by Clifford *et al.*, though was explicitly designed as an abstract mathematical process rather than a model for exploring conflict between species (Clifford and Sudbury, 1973).

Social impact theory considers N agents with three variables - a Boolean attitude o, and real-valued variables persuasiveness p, and supportiveness q (Szamrej et al., 1990). Persuasiveness and supportiveness respectively represent an agent's ability to change the opinion of others, and their ability to influence others to resist having their opinion changed. A property of each edge, immediacy d, describes the ease or probability of communication between any two given nodes. On a lattice this is the Euclidean distance between those nodes, while in other graph topologies it can be handled with weighted edges.

The total conflicting and matching impact, I_c and I_m is calculated using the following equations. In equations 2.1 and 2.2, c and m denote conflicting and matching opinions, respectively. The persuasive and supportive impacts, I_p and I_q of individual agents is scaled by the square of their distance d to the chosen agent, the mean impact found, and then multiplied by the square root of the number of agents performing that impact.

$$I_c = N_c^{0.5} \frac{\sum p_i / d_i^2}{N_c} \tag{2.1}$$

$$I_m = N_m^{0.5} \frac{\sum q_i / d_i^2}{N_m} \tag{2.2}$$

If $I_p > I_s$ then the persuasive impact is greater than the supportive impact, and so the agent is persuaded and changes its opinion. As a neutral ground to observations of real life - a new convert could be impassioned with their new cause and thus more

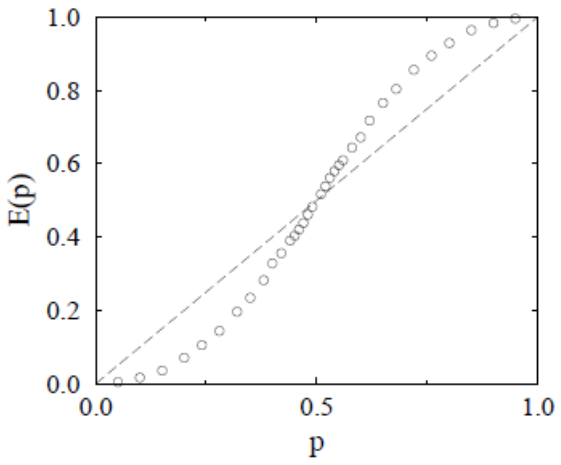


Figure 2.1: Probability that an opinion in a two-state system with initial proportion p dominates. Dots indicate Majority Rule, while the dashed line indicates Voter Model

persuasive, or they could be dismissed as unstable or lacking conviction (Szamrej et al., 1990) - an agent changing its mind update its persuasiveness and supportiveness to random values.

Similarly to the model used by Clifford (Clifford and Sudbury, 1973), the Majority Rule model considers agents with discrete opinions (Krapivsky and Redner, 2003) meeting in groups, to model the decision making processes within companies. However, rather than selecting a pair of agents at each time step, a connected group of more than two agents is chosen. The opinions of all agents within that group are then set to the majority opinion of the group. This more accurately represents the form of decision making seen in meetings, conference calls, and juries. The authors found that for two-state systems in lattices of more than one dimension, the initial majority opinion will usually dominate, according to figure 2.1. However, changes in network topology or number of opinions can give rise to complex dynamics.

The Snazjd model extends the Ising spin model of ferromagnetism to opinion formation (Katarzyna Sznajd-Weron, 2005). Again, a pair of neighbouring agents are considered. If the pair are in agreement, then all neighbours of either member of the pair adopt the opinion. This represents the magnified persuasive power of groups over lone individuals.

If the pair do not agree, then no opinions are changed. Original versions of this model had disagreeing pairs influence their neighbours, but this always leads either to a fully ferromagnetic or fully anti-ferromagnetic state - an unrealistic simplification. This model is based around the trade union maxim "united we stand, divided we fall". An opinion shared by a united group of people will be more easily spread than one held by a lone agent. It references the concept that a group of people have a magnified persuasive power, leveraging people's inherent tendency to want to be part of a group.

2.1.3 Continuous Models

Many of the more recently developed models for opinion formation are continuous models: those that allow an agent to hold an opinion on a continuous scale. This correlates to degree of alignment with a particular viewpoint, rather than a selection from finite choices.

The bounded confidence model was developed by Deffuant et al. (Deffuant, Neau, et al., 2000). It explores opinions as real numbers rather than one of a finite number of options, representing the fact that people express a whole range of opinions on a given topic and can rarely be categorised into agree or disagree so simply. Furthermore, the model enforces that compromise can only be achieved if there is some overlap in opinion to begin with, and that two agents with opinions too far apart will simply refuse to change their opinion on a given topic when conversing.

This continuous model considers N agents and selects a random agent i and one of its neighbours, j each time step. If the difference in the opinions, s, between these agents is within the threshold τ, the opinions of each are adjusted according to the rules in equation 2.3. μ is used as a dampening factor to control the speed at which opinions change, and is equal to or less than 1.

$$s_i = s_i + \mu(s_j - s_i)$$
$$s_j = s_j + \mu(s_i - s_j) \tag{2.3}$$

This model frequently leads to groups of similar beliefs forming, with isolated extremists delimiting the borders of groups. The final number of groups, P, increases as τ decreases, according to $P = 1/2\tau$. This demonstrates that more confident or well-informed agents are more inclined to create smaller clusters around opinions they agree with, rather than drastically alter opinion. N and μ mostly affect the speed of convergence rather than the end results. End results are typically central clustering, convergence to each extreme, or convergence to a single extreme. Increasing the threshold - equivalent to increasing uncertainty - tends towards convergence to extremes.

A significant modification to the bounded confidence model, named Relative Agreement, was developed in 2002, and features individual uncertainty u as the threshold when determining if two agents may interact (Deffuant, Amblard, et al., 2002). This model was devised to investigate how extremists, who by definition hold minority views, can take control of entire networks of moderate agents. Interactions are scaled by h_{ij}, the proportion of u_i that overlaps with u_j. Furthermore, the uncertainty of agents is also modified as a result of interactions - such that interacting with a highly confident individual that you agree with increases your own confidence in that shared belief.

Equations 2.4 through 2.6 and figure 2.2 show how the opinions s are updated according to the degree of overlap between agents i and j.

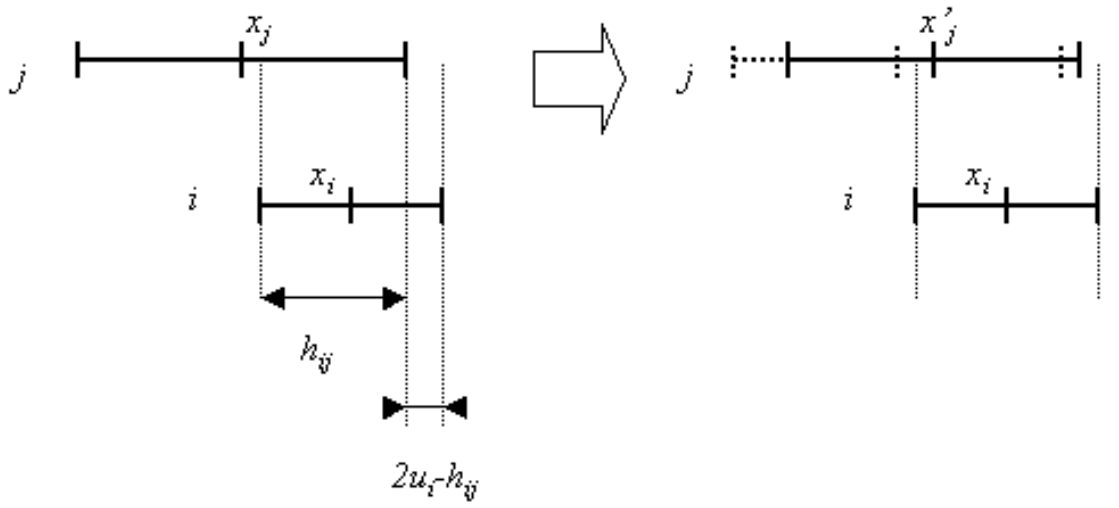


Figure 2.2: The overlap in opinions, or agreement, between agent i and a more confident agent j.

$$h_{ij} = \min(s_i + u_i, s_j + u_j) - \max(s_i - u_i, s_j - u_j) \tag{2.4}$$

$$s_i = s_i + \mu(s_j - s_i)\left(\frac{h_{ij}}{u_i} - 1\right) \tag{2.5}$$

$$s_j = s_j + \mu(s_i - s_j)\left(\frac{h_{ij}}{u_i} - 1\right) \tag{2.6}$$

This model has been used extensively for investigating the role of extremists - agents can be added with an extreme opinion in one direction or the other, and a very high confidence in that opinion, and therefore emulate an extremist who holds a radical view with extremely high conviction. The results of their experiments show the power of the masses over the effects of particular agents, in a significant departure from the standard bounded confidence model. Within this model, three different stable patterns can be observed: either the network drifts to a single extreme, splits to two extremes, or the network is able to resist the influence of extremists and consolidate around a central opinion.

The Continuous Opinions and Discrete Actions (CODA) model describes a situation

in which agents hold a real-valued opinion s, yet may only express themselves in discrete terms A. Martins, 2008. This was designed to emulate systems where expression is limited, such as voting for a particular party or candidate or the up- and down-vote systems available on many social media websites, and to investigate the dynamics that occur therein. Again, at each time step a randomly-selected pair of neighbours i and j are evaluated, and their internal opinions s updated. If s_j is positive, s_i is incremented by a step size α. If it is negative, it is decremented instead.

$$s_i = s_i + h_i \alpha \frac{s_j}{|s_j|} \tag{2.7}$$

As the step size is the same for all interactions, it serves only to control the speed of stabilisation rather than the end result. A heterogeneity factor h_i allows for contrarians and inflexibles to be introduced - those who change their opinion in the opposite direction to usual, or not at all, respectively.

SJBO, or Social Judgement Based Opinions was introduced in 2015 and has a number of features in common with the CODA model from which it draws heavy inspiration (Fan and Pedrycz, 2015). It has two potential scenarios - one in which agents can express their opinion as a real number, and another where they are limited to one of a set of discrete options. In this model, agents are given feedback as to how well a message they send was received, which influences the likelihood of them sending further messages.

Agents have three properties - an opinion s from -1 to $+1$, an assimilation threshold ϵ, and a repulsion threshold τ. At each time step, two agents i and j are selected, and their opinions updated according to the following rules.

$$s_i = \begin{cases} s_i + a_{i,j}(s_j - s_i) & |s_i - s_j| < \epsilon \\ s_i - r_{i,j}(s_j - s_i)\frac{1-|s_i|}{2} & |s_i - s_j| > \tau \\ s_i & \text{otherwise} \end{cases} \tag{2.8}$$

In the scenario with discrete choices, hesitation, h, is added to each agent, and two global properties are added. ρ indicates the decay threshold, below which an agent has not yet become fully committed to a cause, and λ, the decay coefficient. Should the inner opinion of an agent lie within the range of h they will not express an opinion. Otherwise, they will express -1 or $+1$ accordingly. It follows that other agents are unable to discern the true opinion of an agent, only that the magnitude of their support exceeds that of their hesitation. While highly rational agents will believe that agents may not fully support their expressed opinion, to simplify the model agents are believed to fully support their expressed choice. At each time step, a random pair of agents i and j are again selected. If j has expressed an opinion, i updates according to the equations above. Otherwise, if $s_i < \rho$, then $s_i = \lambda s_i$. This reflects an undecided agent who decides to withhold judgement after seeing a lack of conviction among his peers.

This model displays many of the same characteristics as other continuous models, with the addition of hesitation. In a hesitating state the community displays a general preference for one or more options, but with very low consistency. This state either does not stabilise or takes orders of magnitude longer to do so, with agents constantly alternating between expressing either no opinion, and one for which they display a slight preference. This mirrors real-world experience very closely, with agents having no strong feelings either way but a mild preference for one over the other.

Another model was proposed by Nonnenmacher *et al.* in 2014 to investigate how extremism emerges within a society, stating influence from the extreme single polarisation in Germany in the 1930s, and the bi-modal polarisation across the world during the Cold War (Nonnenmacher et al., 2014). In contrast to many other opinion dynamics models, this model simulates N agents moving in a three-dimensional environment, exerting influence on other agents as they move close enough.

Agents have three opinion-related properties: communicability C, influence I, and opinion s, in addition to properties governing their position and velocities. s and C lie

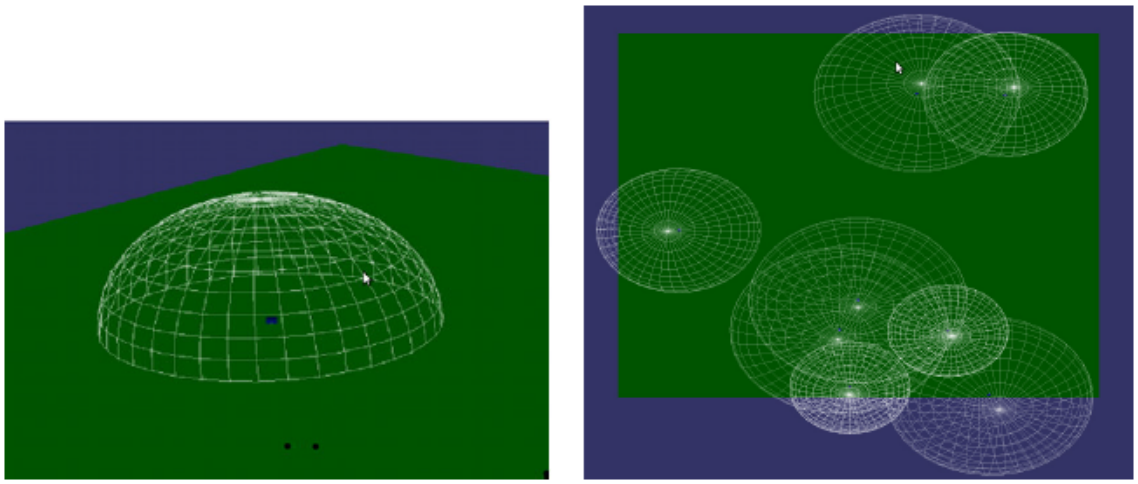


Figure 2.3: Agents, represented by points, are able to influence all those within ranges shown by the wire-frame grids (image from Nonnenmacher et al., 2014)

within the interval $[0, 1]$, while I is unbound. At each time step, each agent i moves across the environment at a speed of C, exerting influence on every agent j that moves within their influence range I according to equation 2.9. A visualisation of this is shown in figure 2.3.

$$s_j = s_j + \frac{I_i(2s_i - 1)}{N} \tag{2.9}$$

This equation gives rise to two groups: those with a state above 0.5 exert a positive influence, while those with a state below 0.5 exert a negative one.

It follows that while increasing both C and I allow a given agent to influence more of its peers, increasing C also allows that agent to in turn be influenced by more agents. Increasing I also has a direct effect on the persuasive power of an agent, whereas increasing C does not. Consequently, the authors find that single extreme convergence is always reached, and that the mean influence I is the most important factor in determining which state dominates.

2.1.4 Other Developments

The effects of external information such as the mass media can be explored within these models. Sznajd-Weron *et al.* explored mass media by creating a virtual agent to represent the media (Sznajd-Weron and Weron, 2008). At each time step, agents had a chance to interact with this agent rather than any of their immediate contacts. This media agent was represented by a group of options, and in each interaction, selected one opinion within that group to portray. This allowed the influence of the media to present different effects to different agents, and across multiple interactions. The authors then explored altering the frequency at which the media presented a given option, corresponding to the idea of paid advertising. Results were then mapped to the Polish telecommunications network market shares, showing close agreement with real-world data (Sznajd-Weron and Weron, 2008).

In this thesis, we focus upon models in which only one opinion is considered at a time. However, multiple-opinion models have been studied, most famously the Axelrod model (Axelrod, 1997). This model defines a culture as a finite group of multiple discrete traits, and explores how cultures can expand, spread, and coexist with one another. In each time step, a random agent i is selected along with a random neighbour j. With a probability equal to the proportion of traits the two agents have in common, i adopts one trait from j. In this way, agents that already have some opinions in common are more likely to adopt further traits. It also follows that agents with nothing in common will never grow closer together unless a third party intervenes. The authors propose a wide array of extensions to this model: representing cultural drift through spontaneous modification of traits, cultural divergence through having a repulsive interaction between sufficiently dissimilar agents, and different network structures such as scale-free, random, and lattices.

2.2 Empirical Studies

To date, there have been very few empirical studies of opinion dynamics (Castellano, Fortunato, and Loreto, 2009; Xia, H. Wang, and Xuan, 2011), owing to the significant practical and ethical issues involved in experimenting upon the opinions of large groups of people. However, a number of studies have observed human behaviour and attempted to fit existing models to the observed patterns, or derive new models from them.

Rouchier and Tubaro used data collected from the Commercial Court of Paris in 2000, 2002, and 2005 to conduct a case study on advice-seeking behaviours among judges (Rouchier and Tubaro, 2011). They employed the relative agreement model, with their strictness or tendency to award punitive damages substituting the opinion. They reasoned that a judge seeking advice from a less strict senior judge would likely become less strict, and the converse. The authors constructed a network from responses from the judges, and then investigated how altering the reliance on authority or position within the hierarchy, reputation, and self-similarity altered the resulting dynamics within the simulation. Of particular note is that judges within the court serve 14-year terms, leading to a regular influx of new (and thus of low authority and reputation) judges and the departure of more senior figures, as well as the gradual increase in authority of remaining members.

The authors suggested three proposals for how agents should form relationships. The authors conclude that when agents form relationships according to authority and reputation, this most closely matches to the observed real-world results (Rouchier and Tubaro, 2011).

Robinson *et al.* used survey data on sexual attitudes and lifestyles to construct a simulated model of a collection of agents, and then employed an SIS model to simulate the transmission of an STD through that network (Robinson, Cohen, and Colijn, 2012). The authors investigate how altering the length of time a patient remains in the infectious

state alters the overall pattern of disease progression throughout the network, finding that infections centre upon a small number of high-activity individuals with a high number of concurrent partners. This link was especially strong when the the duration of infectiousness was particularly low. The authors then investigated a number of possible mitigation strategies such as vaccination and behavioural interventions designed to reduce the spread of disease (Robinson, Cohen, and Colijn, 2012). This demonstration that interventions on particular high-risk individuals are more effective than less targeted interventions ties in neatly with our work, as we look to find if similar interventions can work on non-biological "contagious" phenomena leading to polarisation.

2.3 Limiting the Spread of Biological Infections using Agent-Based Modelling

Work on preventing ideological spread through online social media is limited, due to how relatively new this attack vector is. However, the parallels to medical infection grant us an opportunity to refer to experiments in that field for potential intervention strategies. Exploring what works under these circumstances may allow us to better understand infection in other spheres, such as ideological.

A frequently seen model in work on epidemiology is the SI model, and its variants (Hadjichrysanthou and Sharkey, 2015; Chen et al., 2011; Liu et al., 2016; B. Gao, Deng, and Zhao, 2016). In this model, an agent can be in one of two states - Susceptible, or Infected. Every time step, each infected agent has a chance to infect each adjacent susceptible agent. This model is particularly suited to rumour spreading or incurable diseases: once an agent is infected, they remain so until the end of the simulation.

This family of SI models are a type of discrete model that models potential infection victims as machines with a finite number of states, rules for transitioning between those

states, and rules for triggering transitions in others. For instance, an incurable disease possesses Susceptible and Infected states, a rule permitting a one-way transition from Susceptible to Infected, and a rule allowing an infected agent to potentially infect a neighbour.

Modifications to this model involve adding additional states. The SIS model adds an additional rule allowing an agent to revert from infected to the former susceptible state in a similar manner to the common cold. The SIR model adds a Recovered state, and a rule allowing an agent to transition from Infected to this new Recovered state. This allows an agent to recover and then become immune to the disease, such as chickenpox.

Preventing the spread of infection has been a problem that has plagued mankind for millennia. Up until a few hundred years ago, the only solution was quarantining, with vaccination becoming available from the 18th century. One problem with quarantining and vaccination is identifying which people to quarantine or vaccinate and in what order. Agent-Based Modelling has been adopted to tackle this problem.

Pastor-Satorras and Vespignani model infections using the SIS model on Watts-Strogatz and Barabási–Albert networks, as examples of complex networks with different levels of homogeneity (Pastor-Satorras and Vespignani, 2001). They use three methods of identifying which agents to vaccinate: uniform selection, proportional selection, and targeted selection. Uniform selection randomly selects agents for immunisation, regardless of their connectivity. Proportional selection vaccinates different proportions of agents according to their connectivity meaning that a higher proportion of agents with a high degree are vaccinated. Finally, targeted vaccination vaccinates agents in order of their connectivity. It is shown that while the techniques have similar efficacy on the homogeneous Watts-Strogatz network, targeted vaccination substantially outperforms other techniques on the Barabási–Albert network.

Robinson *et al.* investigate the spreading of sexually transmitted diseases over a simulated network of relationships (Robinson, Cohen, and Colijn, 2012). They use an

SIS model, where it is possible to recover from an infection but doing so does not grant immunity to said infection. The authors propose three intervention methods designed to reduce the size of an outbreak. These methods simulate real-world healthcare procedures, and are designed to be analogous to treatments that reduce the time an agent is infectious for, vaccination, and behavioural modification aimed at reducing the degree of those agents with a degree above a certain threshold. The vaccination procedure is run in two modes. The first randomly selects some proportion of agents to vaccinate, while the second targets agents with a high degree. It is found that targeted interventions substantially outperform random interventions.

A novel method inspired by the single-celled organism *Physarum polycephalum* is proposed by Liu *et al.*, to identify key nodes for vaccination in both real and simulated networks (Liu et al., 2016). The authors employ an SIR model to simulate the spreading of infection, and compare their method favourably to betweenness, closeness, and degree centrality-based targeting methods. To identify key nodes for vaccination, their method examines the growth and spread of a simulated amoeba through the network. Nodes are represented as food sources, with a certain expenditure of energy required to grow across edges. "Tendrils" that lead to rewarding supplies of food grow larger in turn, and the size of the tendril across a given edge gives a quantitative value to that edge.

Nandi *et al.* propose mixed-integer programming and heuristic-based solutions to inhibit the spread of an infection over a network (Nandi and Medal, 2016). The authors suggest four deterministic measures to use as proxies for network vulnerability, defined by paths and connections between infected and healthy agents. Each of these proxies is converted into an optimisation problem, which is approximately solved through mixed-integer programming. This use of proxies allows the authors to avoid computationally expensive simulation, which is critical for large networks and real-time monitoring and blocking of contact. Mixed-integer programming is then used to find the optimal result for minimising each of these proxies, which then directly leads to reducing the size of an

epidemic.

The Min-SEIS-Cluster was proposed by Santiago *et al.* in 2016, and has many similarities with our chosen topic (Santiago, Zunino, and Concatto, 2016). They modify the SIS model to include an additional stage, exposed. In the exposed state, agents are not yet infected and may not spread the infection further, but they will become infected in time. This can represent a disease with an incubation period, or a period of induction between acquiring a new belief and publicly espousing that belief. They also explicitly consider the cost of action both in financial terms and political instability. They propose an optimisation problem to minimise the size of an epidemic at the minimum cost of action. They develop a Monte Carlo-based algorithm to identify the lowest-cost subset of edges that must be severed. However, this solution has a high computational cost, and would not be tractable on graphs with a very high number of nodes.

Concatto *et al.* propose a genetic algorithm to produce solutions to the Min-SEIS-Cluster problem (Concatto et al., 2017). This genetic algorithm compares favourably to previous attempts at the problem, surpassing the effectiveness of the former Monte Carlo-based approach. In their conclusion, they propose using centrality-based measures and other heuristics in order to inform the evolution of their algorithm. This algorithm suffers from a similarly high computational cost, however.

The SI model and its variants are not directly applicable to our work, due to a number of mechanical differences between radicalisation and biological infection. Radicalisation is a process rather than a one-time event, requiring multiple exposures over a period of time to be effective. Additionally, a key concept within opinion dynamics models is the idea of compromise, which provides the possibility of a radical becoming reformed through contact with moderate agents. However, infectious diseases cannot be cured via exposure to healthy agents. We thus turn to a selection of models that allow us more fine-grained control over an agent's opinion, and allow for more precise expression of radicalisation, extremism, and persuasion.

2.4 Limiting the Spread of Extremism

Social media operators are placed in a difficult position when discussion turns towards the prevention of extremist views spreading throughout their platform. On the one hand, there is a benefit to society as a whole when extremism is contained and neutralised, political discussion is non-polarised and fruitful, and collaboration between users of all political perspectives is possible. On the other hand, extreme and radical views increase user engagement and thus directly lead to higher income for the media companies hosting said views. In the case of political partisanship and polarisation there is a further complication, which is that censorship - either real or perceived - of a political figure draws the social media company into direct conflict with the very individuals responsible for passing laws to regulate social media. Unwillingness to be drawn into this conflict has led some social networks including one of the largest at the time of writing, Twitter, to explicitly exempt political figures from their rules and terms of service governing hate speech (Twitter, 2019), and for Facebook to exempt political figures from their rule that advertising must be truthful (BBC News, 2019b).

Social networks have also become a battleground for supporters of and opposition to absolute freedom of expression. Article 19 of the Universal Declaration of Human Rights states that "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers" (*Universal Declaration of Human Rights* 1948), and the first amendment to the United States constitution also supports a view of a right to absolute freedom of expression (*First Amendment* 1791). The International Covenant on Civil and Political Rights qualifies this right, stating that this comes with certain duties and responsibilities necessary for the protection of others, public order, health, security, and morality (*OHCHR | International Covenant on Civil and Political Rights* 1966). Similarly, the global nature of social media

brings social media companies - most of which are based in the US and thus protected by the first amendment - into conflict across the rest of the world, most of which has laws outlawing hate speech. Any attempt to appease foreign governments by complying with their laws may then be seen as a betrayal of their home government, risking considerable backlash.

There is also a problem in how easy it is to recover from having an account deleted. Many of the accounts used to spread extremist propaganda are bots: automated accounts used to create the appearance of a community. While deleting a bot account causes that bot's followers and contacts to be lost, the network of bots is able to quickly regenerate from any small-scale loss. Another way to utilise bots is to select a popular account with many followers likely to be susceptible to propaganda, such as politicians, journalists, or news sites. One of the bots then comments on posts from the targeted account, and the other bots upvote that comment. The comment from the bot is thereby promoted as a top comment and gains visibility from the targeted account despite having very few (human) followers. This allows for the deletion of an account to be almost inconsequential for extremists. There is also the risk that deleting an account triggers backlash from their supporters, such as when conspiracy theorist Alex Jones was removed from mainstream social media (Coaston, 2018).

A related concept dating back to the 1970's is so-called shadow banning (Vice, 2018). An account that has been shadow banned is still able to read messages posted by others and even make their own posts, however their posts are invisible to all other users. This is equivalent to deleting all outgoing edges from a node within a network, without deleting that node itself. This technique is primarily used to combat spam, though it is frequently subject to allegations of being targeted based on political affiliation (Vice, 2018). One drawback to this method is the ease of detection: while harder to detect than a full account deletion, all a user has to do is to log out of their account and then search for their own messages.

There is also a vast amount of information available on social media that is easily obtained. Forcing extremists out of the spotlight and on to private messaging systems like Telegram or to darknet websites may make people of interest that much more difficult to monitor than if they are on the clearnet using services that are compliant to law enforcement requests (BBC News, 2019a).

There has been little work done in preventing influence spread over opinion dynamics models. Work was performed by Kimura *et al.*, who investigate a related problem whereby they attempt to manipulate a network to make it resilient to the emergence of what they call "undesirable influences" (Kimura, Saito, and Motoda, 2009). They introduce metrics for estimating the influence of agents with the aim of blocking the most influential agents so that any undesirable opinions will be unable to spread, regardless of where they first appear. Their ideal solution is too computationally expensive and therefore, they propose heuristics to reduce its execution time. Khalil *et al.* build upon this same influence minimisation problem and demonstrate that finding an optimal solution is NP-hard (Khalil, Dilkina, and Song, 2013). However, they are able to guarantee a solution to their optimisation problem within a narrow margin of the true optimal solution. In our work we do not aim for resilience in advance, but rather efficient intervention once extremist agents have been detected in the network.

2.5 Summary

Opinion dynamics is a wide field, covering many different aspects of human behaviour. Even within the specific topic of extremism, there are multiple approaches and methodologies involved as well as several different models. In chapter 3 we propose a unifying framework that allows us to easily transfer potential solutions between different models without the considerable overhead of needing to develop a new simulator. This framework also allows us to identify shared assumptions, common ground, and differences between

models, as well as determine unexplored regions of parameter space.

It has been shown (Khalil, Dilkina, and Song, 2013) that the influence minimisation problem is NP-hard. Solutions producing near-optimal results have, to date, been computationally complex and take a considerable amount of time to generate results on even relatively small networks. We note that such methods would be intractable upon real-world social networks whose members number in the millions or billions, and so in chapter 4 we propose two algorithms with $O(n)$ complexity that are able to outperform centrality-based algorithms in many circumstances. These algorithms are inspired by previous work across several fields, such as cybersecurity (Mark E J Newman, Forrest, and Balthrop, 2002), healthcare (Robinson, Cohen, and Colijn, 2012), and rumour spreading (Kimura, Saito, and Motoda, 2009), owing to the lack of specific research in extremism prevention. We also draw inspiration from current best practices in de-radicalisation from government sources (Home Office, 2018), which both stresses the need for effective intervention strategies as well as providing direction for our proposed strategies.

Chapters 5, 6, and 6.4 investigate the performance of our algorithms when we remove key assumptions from our model: first the assumption that edges are all equally important within the network, and then the assumption that communications between agents are only ever one-to-one bidirectional conversations. These results represent what happens when agents value different connections to different degrees, and what happens when agents are able to broadcast their opinions via publicly visible posts. We find that factoring cost into our decision making has very little effect on overall efficacy.

Chapter 3

Unified Opinion Dynamics Framework

In this chapter, we propose a framework for opinion dynamics that allows us to decompose existing models into independent parts and then recombine these parts either to replicate existing models or to form entirely new models. Throughout later chapters in this thesis we wish to explore altering methods by which agents communicate as well as the structure of the network while still adhering to the same rules otherwise, following the experimental principle of only changing one variable at a time. As such, we depend heavily on being able to alter different aspects of a model while leaving others untouched and so this framework-based approach is required to allow us to perform these alterations in a systematic and efficient manner. This mirrors the use of replaceable parts in machinery: known-good components can be substituted in to a larger device to alter its functionality without necessitating a rebuild of the entire device.

Castellano *et al.* (Castellano, Fortunato, and Loreto, 2009) observed that "the development of opinion dynamics so far has been uncoordinated and based on individual attempts, where social mechanisms considered reasonable by the authors turned into mathematical rules, without a general shared framework", a sentiment that is later echoed by Xia *et al.* (Xia, H. Wang, and Xuan, 2011) who added that "the related endeavors are largely uncoordinated and presently it may be difficult to construct an

integral and coherent framework for [sic] cover the important aspects of all the related endeavors". In both cases, the authors also mention the interdisciplinary nature of opinion dynamics and how this can lead to confusion. We propose that by standardising terminology and streamlining the simulation of opinion dynamics, experts can easily be brought together across the varied fields of statistical physics, sociology, computer science, and many more to lend their unique expertise. Each of these fields has contributed in their own way throughout the evolution of opinion dynamics, from its early roots in statistical physics, its use of network theory from computer science, and the behavioural insights from sociological studies. We hold that by establishing a common methodology, communication between these fields can be improved and lead to more discoveries.

3.1 Related Work

The Relative Agreement model by Deffuant *et al.* as described in section 2.1.3 was initially proposed for investigating the spread of extremism through a community (Deffuant, Amblard, et al., 2002). The authors found that under many circumstances, the population of the model would adopt a highly polarised state, with almost the entire population becoming either extremely supportive of or extremely opposed to a given idea. This makes it an ideal candidate for our experimentation into the role of polarisation in communities. However, this model is defined as a collection of agents in a complete network where at each time step random agents interact with each other in pairwise conversation, After each interaction the pair of interacting agents update their opinions and confidence based on the given equations in that paper. See section 2.1.3 for more details. However, online social media cannot be thought of as random pair interactions, as there exists a complex graph of who communicates with whom. Thus, we must first alter the Relative Agreement model in such a way as to change the network structure without altering any other part of this network. This shows us that models are not one indivisible part and

that they ought to be thought of as different components that are mostly independent and interchangeable.

We hold that the Relative Agreement model consists of a number of independent rules as follows: the initial structure of the network, the initial distribution of agents' uncertainties and opinions, how agents decide with whom to communicate, and how agents update their uncertainty and opinion as a result of said communications. With this observation we build upon the work by Urbig *et al.*, who proposed the idea of separate regimes that combine to form an opinion dynamics model (Urbig, Lorenz, and Herzberg, 2008). In that paper, the authors successfully recreated the Hegselmann-Krause (HK) model (Hegselmann and Krause, 2002) and the Deffuant-Weisbuch bounded confidence model (DW, or simply Bounded Confidence) (Weisbuch, Deffuant, and Amblard, 2004) by using the same rule for updating opinions, but different rules for deciding with whom to communicate. This work also demonstrated new dynamics using their proposed random-m communication rule. This communication rule unified the HK and DW models into one single model, and also demonstrated unexplored parameter space with new and interesting dynamics that the two earlier models had been unable to investigate.

3.2 Proposed Framework

Following our investigation of work using the models described in sections 2.1.2 and 2.1.3 we revealed four independent modules: structural, communication, update, and co-evolutionary. We consider a module to be a set of related rules governing a single aspect of a model.

Structural Module The module consists of rules that describe the initial population before the simulation begins. They encompass not only the initial distribution and configuration of attributes such as opinion, but also the edges in their network. While

early models were frequently limited to complete graphs, lines, or finite lattices (Clifford and Sudbury, 1973; Katarzyna Sznajd-Weron and Sznajd, 2001), recent modelling techniques allow for scale-free and small-world networks to be investigated as a more realistic model of human relationships. Several models hold global properties like bounded confidence interval. These can be modelled in the framework as a property homogeneous for every agent, and so be assigned at the same time as other structural and initial conditions. Agent attribute proportions such as contrarians and extremists are also seen as structural rules, as the rules governing these proportions only affect the initial composition of the population (Deffuant, Amblard, et al., 2002; André C. R. Martins and Kuba, 2009).

Communication Module Rules in the communication module handle who interacts with whom. Mathematically, they produce a directed subgraph where an edge from A to B means that A is able to communicate with B. The two most common forms of communication rules are *pairwise* and *group*. Pairwise communication selects a random agent and has that agent communicate with a random neighbour. By contrast, group communication selects a random agent and all of its neighbours, and each agent within that group communicates with every other agent in that group. This communication need not be reciprocated: it is possible to communicate a message to an agent without receiving one in return. This fact is used in another form of communication rule used in this thesis, the broadcast group. With this rule, a single agent is a selected to communicate to all of its neighbours in a non-reciprocating manner.

On occasion, agents do not interact with another agent, but rather with an abstract concept (such as mass media, or their own reluctance to speak out loud) or a mathematical ideal such as the average opinion of a group. This can be modelled by creating a temporary "virtual" agent with the desired characteristics and then deleting them after the communication has taken place.

Update Module The update module contains rules governing how agents change or update their opinion as a result of interaction with another agent. These are the central rules in opinion dynamics models, and so update modules are named after the model they originate from.

Once agents i and j are chosen to interact by the communication rules, the update rules determine their resultant changes in opinion. Certain update rules (for instance, Relative Agreement) also alter other parameters about the agents such as their uncertainty. The update rule allows changes such as $s_i = s_j$, to set agent i's opinion to that of agent j, or more nuanced changes such as the gradual shift exhibited in the Relative Agreement model. The equation used within continuous update rules is frequently some variation of $s_i = s_i + \alpha(s_j - s_i)$, where α is some scaling factor.

Co-evolutionary Module Unlike update rules, co-evolutionary rules affect the structure of the graph itself rather than individual opinions. These changes can include adding or removing nodes or agents, and changing, adding, and deleting edges or relationships. The name is borrowed from biology, where the evolution of one species (e.g. prey) is triggered by the evolution of another (e.g. its predators) and that again triggers evolution of the second and so on *ad infinitum*.

Co-evolutionary changes can be any of random, targeted, or reactive. In the random case, elements are altered according to a random selection, such as rewiring an edge after every interaction whether an opinion is updated or not. In the targeted case, they are altered by some selection algorithm such as age-based removal of agents, or based on their opinions. Reactive changes occur as a result of a failure state in the communication rule. For example, if two agents are unable to reach common ground through bounded confidence mechanisms, they may sever that relationship and seek a new agent with which to interact (Kozma and Barrat, 2007).

Exploration of these mechanics is relatively recent, and so a number of models omit

Work	Structural				Communication		Update			Coevolutionary	
	Lattice	Complete	Random	Scale Free	Pair	Group	Majority Rule	Bounded Agreement	SJBO	Null	Edge Deletion
Castellano (Castellano, Loreto, et al., 2005)		✓	✓	✓	✓		✓			✓	
Clifford (Clifford and Sudbury, 1973)	✓				✓		✓			✓	
Deffuant et al. (Deffuant, Amblard, et al., 2002)		✓		✓				✓		✓	
Deffuant et al. (Deffuant, Neau, et al., 2000)		✓			✓			✓		✓	
Fan & Pedryez (Fan and Pedrycz, 2015)	✓	✓			✓				✓	✓	
Fu & Wang (Fu and L. Wang, 2008)	✓				✓		✓				✓
Gil & Zanette (Gil and Zanette, 2006)		✓			✓		✓				✓
Hegselmann & Krause (Hegselmann and Krause, 2002)		✓				✓		✓		✓	
Krapivsky & Redner (Krapivsky and Redner, 2003)	✓					✓	✓			✓	
Urbig et al. (Urbig, Lorenz, and Herzberg, 2008)		✓			✓	✓		✓		✓	

Table 3.1: The independent modules in our framework can be combined to generate opinion dynamics models. Here we present a selection of papers and their constituent modules.

this module entirely. This is modelled as a null rule, which does nothing.

3.3 Reconstructing Existing Models using the Framework

In this section, we describe how the models previously identified within our literature review fit within our framework, and detail the transformations that could undertaken to link very similar models together. Table 3.1 displays the framework components comprising the models used in a selection of papers. Only the Relative Agreement and CODA models have been replicated within this thesis: the remaining models are provided as a starting point for future work and an example of the capabilities of the framework.

Voter Model and Majority Rule Both of these models operate in the same manner - a point in opinion-space is determined through taking the modal average of the group of influencing agents, and the influenced agent moves to that point. The only difference between these two models is the number of agents that are influencing at one time. We express this with an update rule that considers an influenced group G_i and an influencing group G_j, and performs $s_i = \text{mode}(G_j)$ for each agent i in G_i. In the voter model, the communication rule is such that G_i and G_j each consist of a single different agent,

whereas in the majority rule model $G_i = G_j$, and the group size is usually larger. Using $|G| \geq N$ ensures each agent interacts with all its neighbours. These models often use complete networks for structural rules, and a null coevolutionary rule.

Social Impact Theory In social impact theory, the communication rule is that an agent is considered along with its neighbours. The update rule is shown in the equations below, where c and m denote conflicting and matching opinions, respectively. The persuasive and supportive influence of the chosen agent's neighbours is totalled, and opinion updated accordingly.

$$I_c = |G_c|^g \frac{\sum_j^{G_c} p_j / d_j^2}{|G_c|} \tag{3.1}$$

$$I_m = |G_m|^g \frac{\sum_j^{G_m} q_j / d_j^2}{|G_m|} \tag{3.2}$$

Where q denotes supportiveness, p denotes persuasiveness, and g is a factor denoting the relative persuasive power of groups. This factor reflects the fact that the persuasive power of groups does not linearly increase with the size of a group that holds the same opinion.

If $I_c > I_m$ then the agent is persuaded, and we perform the following operation:

$$s_i = -s_i \tag{3.3}$$

Bounded Confidence, Hegselmann-Krause, and Relative Agreement These models consider N agents and for its communication rule, selects a random agent i and one of its neighbours j each time step. If the difference in opinions between these agents is within the threshold d, the opinions of both are adjusted according to the update rule below. As relative agreement is an expansion to bounded confidence, the equations for both are similar. Below, the update rule for bounded confidence is shown

in equation 3.4, and relative agreement in equations 3.5 and 3.6.

$$s_i = \begin{cases} s_i + \mu(s_j - s_i) & |s_i - s_j| < u \\ s_i & \text{otherwise} \end{cases} \tag{3.4}$$

$$h_{ij} = \min\{s_i + u_i, s_j + u_j\} - \max\{s_i - u_i, s_j - u_j\} \tag{3.5}$$

$$s_i = s_i + \mu\left(\frac{h_{ij}}{u_i} - 1\right)(s_j - s_i) \tag{3.6}$$

These can be unified with:

$$s_i = \begin{cases} s_i + \mu\alpha(s_j - s_i) & v(s_i, s_j) = 1 \\ s_i & \text{otherwise} \end{cases} \tag{3.7}$$

where α is some scaling factor and v is a function returning 1 (or true) if an interaction is possible between agents i and j, and 0 otherwise. In order to recreate the bounded confidence model, let $\alpha = 1$ and $v(s_i, s_j) = |s_i - s_j| < u$. To instead reproduce the relative agreement model, let $\alpha = \left(\frac{h_{ij}}{u_i} - 1\right)$ and $v(s_i, s_j) = \text{true}$.

To create the Hegselmann-Krause model, instead of having agent i interact with a random neighbour, create a virtual agent with their opinion set to the mean of all neighbours of i. Using a subset of m neighbours instead of all neighbours results in the random-m model described by Urbig *et al.* (Urbig, Lorenz, and Herzberg, 2008).

CODA The Continuous Opinions, Discrete Actions model (see section 2.1.3) simulates a situation where opinions are held as a real number, but the expression of those opinions is limited to a finite set. This can be used to model reactions on social media such as Reddit in terms of up-voting or down-voting a particular post, which expresses approval and disapproval respectively. However, the intensity of such a reaction is not known.

It can also be used to model voting in representative democratic systems where the infinite possibilities of political opinion must be expressed as a choice from a finite list of candidates.

The CODA model is explicitly acknowledged as an update rule - Martins describes his work as combining the CODA model with the Sznajd model (A. Martins, 2008). By decomposing the resultant work into distinct communication and update mechanisms, we are left with rule modules for Snzajd and CODA.

SJBO The Social Judgement Based Opinion model (see section 2.1.3) considers not only an agents opinion, but also ranges around that opinion that they consider attractive or repulsive. At each time step, two agents i and j are selected using a reciprocating pairwise communication rule, and their opinions updated according to the following update rule.

$$s_i = \begin{cases} s_i + a_{i,j}(s_j - s_i) & |s_i - s_j| < \epsilon \\ s_i - r_{i,j}(s_j - s_i)\frac{1-|s_i|}{2} & |s_i - s_j| > \tau \\ s_i & \text{otherwise} \end{cases} \tag{3.8}$$

Where ϵ is the agent's assimilation threshold, τ is their repulsion threshold, a is the assimilation coefficient, and r is the repulsion coefficient.

3.4 Implementing the Framework

In this section, we describe our implementation of the framework in a simulator. Using this simulator, we replicate prior work which we discuss in the next section.

The simulator is a Python program for constructing and evaluating opinion dynamics models according to our unified framework. It is designed to be easily extensible, system-independent, and simple to use. Several rules are provided for the user, allowing them to begin experimentation without needing to write any code. These rules are all fully

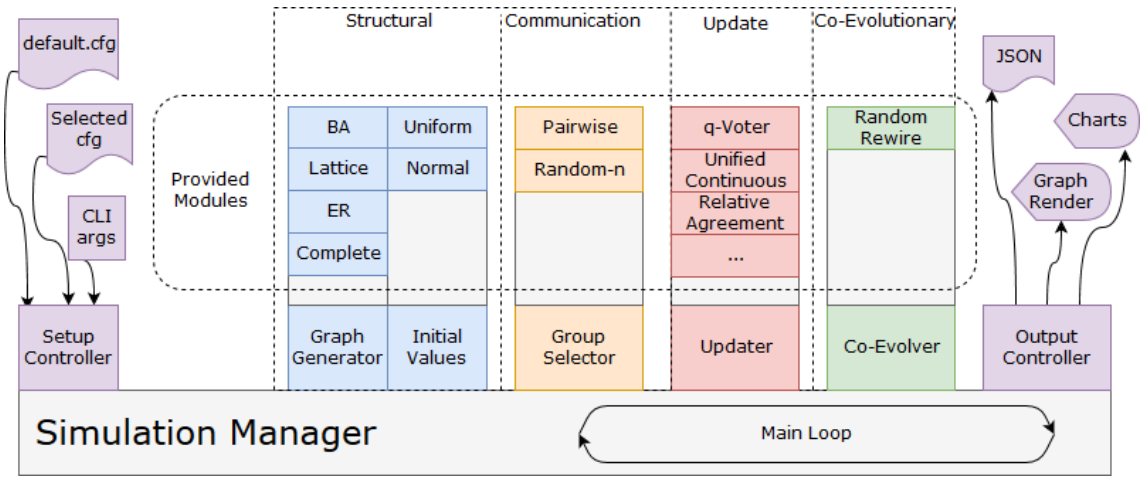


Figure 3.1: Program Control Flow
Control flows from left to right before entering the main loop.
Models are assembled by selecting one module from each column. Some rule options are omitted, for space.

independent of each other, for easy exploration of new combinations of modules. Figure 3.1 shows the control flow of the simulator and how the rules map to the rule components of the unified framework.

The program is invoked from the command line, with arguments controlling which modules to load and the parameters to be used therein. Two configuration files are also used: `default.cfg`, and optionally another `.cfg` file to be provided by the user. If an argument is set in multiple locations, command line arguments take priority over user-specified `.cfg` files, which take priority over `default.cfg`. This allows for easy automation of testing. Throughout this thesis, `default.cfg` was configured with a general set of parameters for the Relative Agreement model and each experiment given its own user-specified configuration file. Tests were controlled using command line parameters provided by the high-performance computing (HPC) cluster's job scheduling system.

Output is then fed to a user-selected output controller. Controllers are provided for rendering the social graph and slowing processing down so that the evolution of opinions can be seen, outputting the opinions of agents over time in a line chart, or saving the

results as a JSON or CSV file for analysis in other programs.

The main simulator modules are briefly described here. One module in each category is dynamically loaded according to the specifications in the configuration files and command parameters.

Graph Generators The complete graph generator produces a graph in which every agent is connected to every other agent. This is the generator used in models where every agent is able to interact with every other agent, such as random pair interactions.

Erdős–Rényi (ER) graphs are a type of graph in which each potential edge has an equal chance to exist (Erdős and Rényi, 1960). This generator takes a number of nodes and a probability and returns a graph with the requested number of nodes, and every possible edge having that probability of existing. This is also simply referred to as simply a "random" graph.

The 2D Lattice generator produces a square graph of n agents in which each agent is connected to a neighbour to the north, south, east, and west, save for those agents on the edge or in the corners of the graph.

Barabási–Albert (BA) graphs construct a scale-free network using a preferential attachment model, in which edges are added from new nodes to nodes that already have many edges. Scale-free networks obey a power law in their degree distributions: a few agents have orders of magnitude more relationships than others.

The small-world generator produces a graph in which the length of the path between any pair of agents grows proportionally to the logarithm of the number of agents in the network.

These graph generators are all available as part of the Networkx Python module, used in our simulator (Hagberg, Schult, and Swart, 2008).

Initial Values The Uniform Distribution generator returns a number uniformly selected from the interval between -1 and $+1$, inclusively.

When provided with a mean and standard deviation, the Normal Distribution generator returns a value selected from that normal distribution, capped to within -1 and $+1$. This generator defaults to $\sigma = 1$ and $\mu = 0$, the standard normal distribution.

Group Selectors The Random-n group selection module uses the selection process of Urbig *et al.* (Urbig, Lorenz, and Herzberg, 2008). Given an agent, it returns n neighbours of that agent. In addition, the Pairwise module is provided as a convenient shorthand for Random-n with $n = 2$. If n is greater than the number of neighbours, all neighbours are returned. Consequently, a Group module is provided as shorthand for Random-n with $n = \infty$.

Updaters The two main update modules are capable of emulating many of the existing models. In particular, the q-Voter model within the framework emulates discrete models using a variant of the voter model, and the Unified Continuous model (see section 3.3) produces the continuous models using a variant of bounded confidence.

For convenience and ease of replication of earlier work, modules are provided for those models covered earlier. For instance, the relative agreement model provides a function for calculating the overlap of two agents' opinions to the Unified Continuous module, and then returns the result of that module.

Co-Evolver The Random Rewire co-evolutionary module takes a pair of agents, severs the edge between them, and then establishes a new edge to a randomly chosen node in the graph.

As many models do not use co-evolutionary rules, we have also provided a Null module. This simply performs no alteration to the graph and then returns.

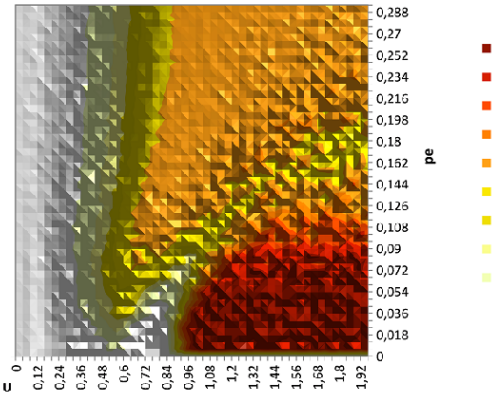
Once the desired modules have been loaded and initialised, the rules in the structural module are executed to create a network of agents and the program then enters its main loop. In this loop, group selector modules choose agents to interact, and updater and co-evolver modules update agents and the network as a result of these interactions. Results are then sent to an analysis module for processing and output in the form of comma-separated values, JSON, or rendered as images.

3.5 Validating the Framework

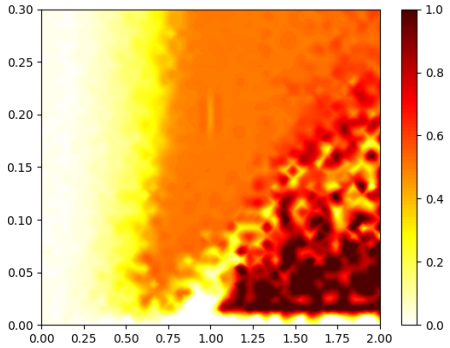
Opinion dynamics models rely on simulated data, and as shown in section 3.5.1, are highly sensitive to small changes in the model. We validate our work by recreating simulations and producing near-identical results to existing published work. Original simulators were homogeneous and fully integrated: simulation software was purpose-built for each experiment. Using our framework, we break models down into independent modules, yet produce the same the same output for a given set of parameters as the purpose-built simulators. Despite the conceptual and programmatic changes between our implementation and theirs, a faithful replication of the output demonstrates that the framework is valid, that the models are composed of fully separable modules and can be perfectly recreated with the independent modules of the framework.

3.5.1 Relative Agreement

We selected two pieces of work from Deffuant *et al.* to compare against (Deffuant, Amblard, et al., 2002; Deffuant, Weisbuch, et al., 2013). In this second piece of work, the authors contest findings by Meadows and Cliff (Meadows and Cliff, 2012) that differed from their own work. Meadows and Cliff had published different findings using the same parameters as Deffuant *et al.*, and used these findings to propose alternative conclusions to the question of how extremism spreads through networks.



(a) A heatmap of the y metric from the Relative Agreement model using the corrected, purpose-built simulator from Meadows and Cliff(Deffuant, Weisbuch, et al., 2013).



(b) Using our framework to implement the Relative Agreement model by linking independent modules reproduces the heatmap almost exactly.

Figure 3.2: A replication of the heatmaps generated by the Relative Agreement model.

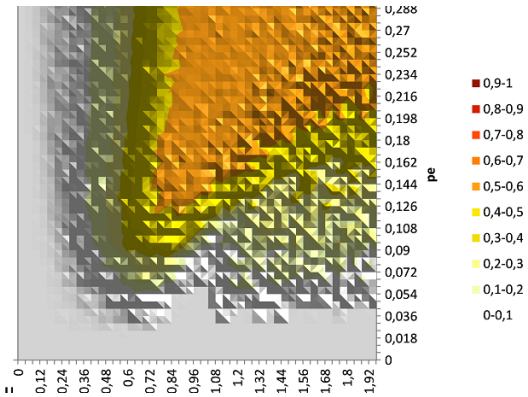


Figure 3.3: Using a simulator with two minor discrepancies from the original model produces a significantly different heatmap of the final metric.

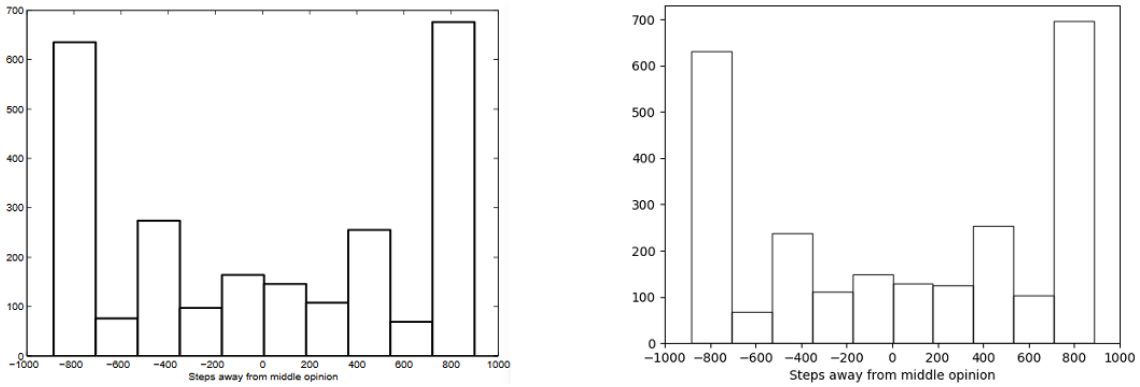
In the initial study by Meadows and Cliff (Meadows and Cliff, 2012), a minor programming discrepancy that partially arose from the model not being fully specified led to vastly different results than were expected, and they were unable to replicate the findings of Deffuant *et al.*. Firstly, the simulation was not carried out for a sufficient number of time steps, and so metrics were calculated before the model had fully converged. Secondly, the threshold to be considered extreme was intended to be lower at the end of the simulation than at the beginning, leading to a far smaller number of extremists being reported under certain circumstances. Thankfully, Deffuant *et al.* were available to consult, and after resolving these issues both groups arrived at very similar results.

This offers a unique opportunity to compare our work with multiple authors performing the same simulation, each using different purpose-built, one-off programs. These experiments made use of the relative agreement model in random pair interactions on a complete graph. In each simulation, the proportion of extremists p_e and the global uncertainty of the non-extremist agents U were varied, and the resultant trends plotted in a heat map, in figure 3.2a. To analyse this trend, Deffuant introduced a metric, $y = {p_+}^2 + {p_-}^2$. p_+ and p_- represent, respectively, the number of initially moderate agents who became positive extremists, and those who became negative extremists. A y value of 0 indicates no polarisation, a value of 0.5 indicates extreme polarisation, and a value of 1 indicates that every agent converged to the same extremist side - either $+1$ or -1.

Our primary validation method is through visual comparison of a graph produced by our study with the graphs produced by the two aforementioned studies, aiming to identify common features in each set of results. Inspection of figure 3.2 reveals a number of artefacts:

1. A white section on the far left, of central or no convergence.

2. A large red area in the bottom right, of extreme single convergence.



(a) Martins' simulator **(b)** Our simulator

Figure 3.4: The distribution of final opinions in terms of the number of steps away from 0, after 800 runs of the CODA Voter update rule.

3. An orange wedge, from the top right towards the centre, of extreme dual convergence.

4. An unusually low-valued line separating the red and orange areas.

This test showcases the importance of small factors in this form of simulation. As two relatively minor changes were able to introduce drastically different results to those seen in the original work and the corrected work by Meadows *et al.*, we see that the model is highly susceptible to minor changes (compare the erroneous heatmap in Figure 3.3 to the correct one in Figure 3.2a. Thus, our simulator producing a graph very similar to the valid one (see figure 3.2b) demonstrates that the framework is valid and that it has been implemented correctly.

3.5.2 CODA

For a further test, we then use the simulator to replicate the results of Martins (A. Martins, 2008), seen in figure 3.4. We do this by analysing the results of our simulator after swapping two rules. The structural rule is changed from complete to lattice, and the update rule is changed from Relative Agreement to CODA.

In the CODA model, described in section 2.1.3, agents select a random neighbour and update their opinion according to the following rule. If the neighbour's opinion is greater than 0, they increment their opinion by a constant amount called the step size. If it is lower than 0, they decrement their opinion by that amount. We reconstruct the author's purpose built model in our framework by using the 2D Lattice graph generator, the Uniform initial value module, the Pairwise (or Random-2) group selector, and an update rule to handle the previously-mentioned rule. 2500 agents in a square 2D lattice are seeded with initial opinions between -1 and $+1$, inclusive, and a step size of 0.2, and then left to interact 800 times each according to the CODA update rule.

We measure the distance from 0 of each agent's opinion after $k = 800$ iterations per agent, in terms of multiples of the step size. The histogram displays a penta-modal distribution with primary peaks at $x = k$ and $x = -k$, secondary peaks at $x = k/2$ and $x = -k/2$, and a tertiary peak at 0. This demonstrates that for a 2-dimensional lattice, the CODA update rule results in most agents adopting a position as far from zero as possible, approximately half that far, or a position of zero. The strong similarity demonstrates that this model also fits the framework.

3.6 Conclusion

The framework presented in this chapter shows a structural similarity existing in many of the opinion dynamics models we have identified. As shown by Urbig *et al.*, exploration of individual modules within the framework can yield interesting and novel results (Urbig, Lorenz, and Herzberg, 2008). We can also reduce or eliminate the possibility of errors in replicating models through verified, publicly available code.

In the remainder of this thesis, we propose intervention methods that aim to reduce polarisation in the Relative Agreement model. We use our software to explore the Relative Agreement model in various circumstances, changing the structural, communication, and

update rules independently and in conjunction with one another in order to understand better how it reacts to our proposed interventions.

The use of simulated data is a central feature of opinion dynamics, allowing us to simulate scenarios that would be impossible to study in real life. In constructing and verifying the framework, we have made use solely of simulated data sets using the same methods as the original authors to which we compare our work, and models that are considerably abstracted from human interactions. In future work we would like to explore the possibility of investigating real-world interactions and data sets to see if such a clear delineation exists between how we choose with whom to interact, and the outcome of those interactions. Most notably, real-world interactions are rarely solely pairwise or solely group-based, but rather we experience a mixture of these types of interactions throughout the day. These strict separation of duties between group selector and updater also misses the fact that many of us behave differently in groups than we do in private conversations. However, the lack of longitudinal studies using real-world data in opinion dynamics offers us very limited opportunities to validate our models against.

Chapter 4

Intervention Algorithms to Reduce Network Polarisation

Within this thesis, we are attempting to prevent or reduce polarisation of a network of interacting agents by means of altering the structure of that network. We have a further goal of keeping these changes as non-invasive as possible, and so we only consider the removal of edges, rather than nodes.

As the novel contribution for this chapter, we propose that non-invasive or light-touch interventions can increase the likelihood of compliance by social media companies, be perceived more favourably by their users in both absolute and conditional free speech countries, have a lasting effect less easily circumvented than account deletion, and leave targets available for information gathering. To this end, we initially defined 8 algorithms for intervention, that each list edges in an existing network in a priority order for removal to reduce network polarisation. This simulates a third party such as network moderators or governmental figures intervening on the network. We run an opinion dynamics model on a network, and can compare the results of this model with and without an intervention algorithm. These were categorised as either opinion-agnostic, or opinion-aware. While both types require knowledge of the network structure as well as whether each agent is

extreme or not, opinion-aware algorithms use information about the exact real-valued opinion of each agent. This allows us a finer granularity of information, such as knowledge of whether an agent is in close agreement with its neighbours, and whether an agent is close to becoming an extremist. Taking into account this greater degree of information may allow our intervention algorithms to perform better.

To be able to gather this information is a difficult problem and thus far not fully solved. A variety of solutions have been proposed over the years. Alizadeh obtains a list of extremist organisations from the US Department of Homeland Security and classifies Twitter followers of those organisations as extremists, after taking steps to eliminate journalists, researchers, and other interested non-extreme parties (Alizadeh, 2012). These steps include setting an upper and lower bound on follower count, eliminating verified accounts, and accounts following both left-wing and right-wing extremists simultaneously. Linguistic analysis-based methods have also been proposed (Torregrosa and Panizo, 2018) that take account of various keywords to estimate the level of radicalisation of a specific account. However, this approach is prone to error: an attempt by a group of hackers affiliated with the Anonymous movement to classify extremists wrongly categorised figures such as BBC News and Barack Obama as supporters of Islamic State (BBC News, 2019a). Additional work has expanded the linguistic approach by taking context into account (Fernandez and Alani, 2018), and models of social influence from the social sciences (Fernandez, Asif, and Alani, 2018).

In the experiments in this thesis we have abstracted away the process of obtaining information with such precision.

4.1 Intervention Algorithms

In this section, we present the algorithms that we initially tested to determine centrality, which we use as a proxy for importance within the network. In order to increase the

chances of compliance by social media companies, we restrict ourselves to so-called light-touch interventions. We avoid interventions that involve removing agents from the network, and instead only remove edges. While removing all edges from an agent is functionally identical to removing the agent itself, each edge removed makes the agent a less desirable target and thus less likely to be targeted by subsequent interventions. We also limit ourselves to removing no more than 10% of the edges in the network in total.

4.1.1 Opinion-Agnostic Algorithms

Opinion-agnostic algorithms use standard centrality metrics to generate a score for each edge, influenced by work on the SI models that took this approach. Where the centrality algorithms generate scores for nodes, the score for an edge is given as the average of the nodes at each end of that edge. The highest-scoring edge leading from an extremist to a non-extremist is then severed. In complexity analyses, V is the set of vertices or nodes in the network, and E is the set of edges. This family of algorithms depend on knowledge of extremist status, but only uses this information to qualify who is and is not a valid target. By contrast, opinion-aware algorithms use this information as part of determining how valuable a given edge is as well.

We follow the example of Newman *et al.* and select random interventions as a control case, and degree based intervention as an example centrality-based algorithm. We also selected three other centrality measurements for initial experimentation, to see if they gave substantially different results and to investigate the execution time of algorithms with a computational complexity class higher than O(n). These measures were included in the NetworkX package, and we theorised that their initial complexity could perhaps offer initial efficacy.

Random The random intervention algorithm selects an agent at random, and severs an edge leading to one of their neighbours, again chosen at random. This serves as a

control case, and a point of comparison for other methods. This method, rather than simply selecting an edge at random from all edges, was chosen for similarity to our other opinion-agnostic methods which all select the highest scoring node, then delete the highest scoring edge attached to that node.

Degree The degree centrality of a node is the number of edges connected to that node. With the exception of random, this is the the least complex algorithm, with a computational complexity of $O(|V|)$. As we are using undirected graphs, there is no difference between total degree as used here, and out-degree as used by Newman *et al.*.

Current-Flow Also known as information centrality, this algorithm envisions a network of electrical resistors, where nodes represent junctions and edges represent resistors. The algorithm then calculates the current flow between each pair of nodes in the network, and assigns each node a score based upon the amount of the total current that passes through that node (Brandes and Fleischer, 2005).

Betweenness Betweenness centrality calculates the shortest path between every pair of nodes within the network. The score of a given edge is the number of those shortest paths that the selected edge appears on. The computational cost of this algorithm is $O(|V||E|)$. This algorithm was chosen as it is a frequently used general measure of centrality, in which a node with high betweenness centrality would have a high level of control over the network.

Eigenvector The Eigenvector centrality of a node is a measure of influence within a network. Similar to the ranking systems used by many search engines, connections from highly-ranked nodes contribute highly to your own rank (Bonacich, 1987).

4.1.2 Proposal: Opinion-Aware Algorithms

In contrast to the purely centrality-based methods, these new algorithms use more information in calculating the importance of a node than the network structure. While opinion-agnostic algorithms only use information about extremist status to discern who is an eligible target, in opinion-aware algorithms information about extremist status and also precise opinion values is used to allow for more precise targeting and selection of edges for removal. In opinion formation models, we have *a priori* information about the state of individual agents within the network, in addition to full information on the structure of the network. Making use of this information allows us to use these novel methods to more precisely target edges for removal, while keeping our interventions lightweight reduces the risk of our targets noticing and responding to our actions and thus requiring more potentially time-consuming information gathering efforts.

We propose a modification to the Betweenness algorithm as well as two new contributions that are all considered opinion-aware algorithms.

Subset Betweenness A variant of Betweenness, this algorithm only considers paths that connect an extremist node to a non-extremist node, on the basis that influence from extremist to extremist is unimportant for our purposes, and influence between two non-extreme agents is acceptable communication.

Influence Targeting Influence Targeting is an algorithm we propose that is a modification to degree centrality. Rather than considering degree alone, we consider the risk posed by an agent being radicalised in terms of the number of others exposed, and also the ease with which this can be prevented. We are then able to divide the "payout" of protecting an agent in terms of polarising influence minimisation by the "cost" of performing interventions to achieve this. Each non-extreme agent's degree is calculated, and then shared evenly among its edges that lead to extremists. For example, an agent

with 12 neighbours, three of whom are extremists, has each of those edges leading to extremists scoring $12/3 = 4$. The edge with the highest score is then severed. Full code for the module in our simulator is shown in listing 4.1.

This algorithm attempts to make best use of limited resources by prioritising those individuals whose radicalisation poses a great risk to the community, by taking into account the degree centrality of that agent. Degree centrality in this case being a useful proxy for influence, this allows us to identify agents who may cause a disproportionate number of agents to become susceptible should they become radicalised. However, it does not attempt to consider the likelihood of this occurring unlike our other algorithm, Vulnerability Targeting.

Vulnerability Targeting We also propose the Vulnerability Targeting algorithm. This algorithm uses information about the precise opinions of agents in order to prefer interventions on agents that are dangerously close to becoming radicalised. In this algorithm, edges connecting extreme agents to non-extreme agents are weighted according to the reciprocal of the distance between the two opinions. This results in a very high weight for edges between almost radicalised agents and their would-be recruiters, and low weights for edges from agents that would require many interactions before succumbing to radicalisation. Python code is shown in listing 4.2. Unlike our two previous algorithms, this requires considerably more information: precise information on the exact opinion value of each agent. This is because rather than just identifying extreme agents, we aim to identify vulnerable agents who are almost extreme.

Much like the UK's PREVENT strategy (Home Office, 2018), this algorithm aims to protect a broader society through protection of individuals. By preventing a vulnerable person from becoming radicalised, it is hoped that not only will that individual be protected from harm, but that their wider community will be insulated from radicalisation. Unlike the Influence Targeting algorithm however, it does not consider the size of that

Listing 4.1: Code for the Influence Targeting algorithm

```
from networkx import Graph, connected_components

def intervene(G, opts):
  # make a temporary copy of the giant component
  temp = Graph(G)
  ccomps = connected_components(G)
  cc = list(max(ccomps, key=len))
  ncc = [n for n in list(G.nodes()) if n not in cc]
  temp.remove_nodes_from(ncc)

  # get a dictionary of "is a node extremist"
  eh = opts.initial.max * 0.9
  el = opts.initial.min * 0.9
  extremists = {
    n: not el < G.values["opinion"][n] < eh
    for n in temp.nodes()
  }

  # assign weights to edges
  targets = {}
  for node in temp.nodes():
    # we don't trim edges from the extremists
    # but rather from their targets
    if extremists[node]:
      continue
    nbors = list(temp[node])
    ex_nbors = [n for n in nbors if extremists[n]]
        # weight is degree, shared
        # between edges to extremists
        if len(ex_nbors):
          weight = len(nbors)/len(ex_nbors)
          for ex_n in ex_nbors:
                targets[(node, ex_n)] = weight

  # remove the edge with the highest weight
  targets = targets.items()
  targets = sorted(targets, key=lambda e: e[1])
  if len(targets):
        edge = targets[-1][0]
    G.remove_edge(*edge)
  return
```

wider community, only the likelihood of extremism occurring within that individual.

The two algorithms we propose are deliberately at either end of a continuum: Vulnerability Targeting considers individual risk without respect to that individual's community, while Influence Targeting only considers the risk to the wider community should an individual be radicalised. We note that in real-world planning one would consider both the likelihood of an event along with the consequences should that event occur, however for the purposes of our experimentation we elect to consider only these two extremes. In future work it would be useful to explore a combination of these factors, and explore both the probability of an individual being radicalised along with the consequences of them being so.

4.2 Experimental Setup

For our experimentation, we use the software described in (Coates, Han, and Kleerekoper, 2018a), (Coates, Han, and Kleerekoper, 2018b), and section 3, with some minor adaptations and a wrapper program to allow it to run on our HPC cluster's job submission system. The HPC is comprised of 12 machines, each with 2 8-core Intel Xeon E5-2650 v2s at 2.60GHz, and 64GiB of RAM.

These adaptations only alter the input/output capabilities of the program, and do not affect the actual simulation process. The wrapper program interacts with the cluster management system to make maximal use of available resources, again without affecting the underlying simulation. We developed this program to demonstrate the validity of our framework in chapter 3. It assembles an opinion formation model from independent modules each governing a single aspect of the simulation: network structure, initial values of agents, how the agents interact, and what occurs as a result of that interaction. It takes input from configuration files such as listing 4.3 or the command line, and can output data in textual format, graphs, or videos showing the evolution of opinion over

Listing 4.2: Code for the Vulnerability Targeting algorithm

```
from networkx import Graph, connected_components

def intervene(G,opts):
  # make a temporary copy of the giant component
  temp = Graph(G)
  ccomps = connected_components(G)
  cc = list(max(ccomps, key=len))
  ncc = [n for n in list(G.nodes()) if n not in cc]
  temp.remove_nodes_from(ncc)

  # get a dictionary of "is a node extremist"
  eh = opts.initial.max * 0.9
  el = opts.initial.min * 0.9
  extremists = {
    n: not el < G.values["opinion"][n] < eh
    for n in temp.nodes()
  }

  # assign weights to edges
  target_edges = {}
  for node in temp.nodes():
        # we don't trim edges from the extremists
        # but rather from their targets
    if extremists[node]:
      continue
    nbors = list(temp[node])
    myop = G.values["opinion"][node]
    ex_nbors = [n for n in nbors if extremists[n]]
    # weight is
    # 1/(distance to closest extremist neighbour)
    if len(ex_nbors):
      dists = [
        myop-G.values["opinion"][n]
        for n in ex_nbors
      ]
      weight = max(abs(1/(n)) for n in dists)
      for ex_n in ex_nbors:
        target_edges[(node,ex_n)] = weight

  # remove the edge with the highest weight
  targets = targets.items()
  targets = sorted(targets, key=lambda e: e[1])
  if len(targets):
    edge = targets[-1][0]
    G.remove_edge(*edge)
```

Listing 4.3: The configuration file used for our initial tests

```
[graph]
n = 200
alg = complete
e = 2

[initial]
uncertainty = 1.0
extremists = 0.1

[group]
alg = pairwise

[update]
alg = relative_agreement
mu = 0.2
iterations_each = 800
```

time. In this case, we output data into .csv files, which are then collated, downloaded from the cluster, and processed locally using the matplotlib and seaborn libraries for Python.

We provide the configuration file in listing 4.3 to the program, with further intervention algorithm details passed via command-line arguments. Each experiment is repeated 128 times: the largest job size that could be run on the shared cluster.

For each experiment, we generate a complete network with 200 nodes. These agents were then seeded with initial opinions uniformly distributed between -1 and 1. Agents that were selected to be extreme had their uncertainty set to 0.1 and their opinion set either to -1 or 1, with both being equally likely. These parameters were select to match those used by Deffuant in his original paper on the Relative Agreement model (Deffuant, Amblard, et al., 2002), ensuring that our results would be comparable and that any discrepancies would be immediately obvious as a flaw in our implementation of the model.

We then apply our intervention algorithms with a set budget of removals. As we are aiming for light-touch interventions, we cap our budget at 10% of the edges removed.

The Relative Agreement model is then run for 800 iterations per agent. With a μ value of 0.2, this should safely ensure that the model has stabilised before we take a final reading of results, as per the recommendations in Deffuant's 2013 work (Deffuant, Weisbuch, et al., 2013).

We initially experiment upon a complete graph, in order to closely mirror Deffuant *et al.*'s original work (Deffuant, Amblard, et al., 2002). This ensures that our results are generated using an exact replica of the Relative Agreement model as defined in that work, before we begin altering different components using the framework described in chapter 3. This experiment, like Deffuant's before it, represents an environment in which every agent interacts with every other agent at random, and is the most simple network structure we can explore. We envision that under these circumstances extremism is all but guaranteed: a light-touch intervention will be insufficient in preventing enough communication with extremist agents to prevent the drift to extremism, and thus network polarisation.

4.3 Test Cases

Before running our simulations we proposed that work on a complete graph would be a useful test of our simulator's capabilities. Given that a light-touch intervention would not substantially reduce the overall connectivity of the graph, we theorised that said light-touch interventions would not have much effect on the spread of extremism or polarisation in this scenario.

Test Case 1 *There will be very little difference between the effect of zero interventions and a relatively large number of interventions (i.e. 3000) on a complete network.*

After work on a complete network, we move onto the Erdős–Rényi random $G(n, p)$ network. This is a well-studied and understood network structure that offers us a stepping stone towards the types of network structure exhibited by real-world social networks

(Erdős and Rényi, 1960). In the complete network, before we perform any interventions the structure is essentially homogeneous: every agent has the same options available to it, that is, to interact with anybody else. Given that all of our interventions depend to some extent on the network structure, this is undesirable. This also means that no extremist is more influential than any other, since all have the same reach. In a random network however, the structure of the network limits the options available to each agent, and so leads to some extremists having more influence than others. On this network we expect to find that opinion-aware algorithms tend to have superior efficacy to opinion-agnostic algorithms, especially in conditions of low uncertainty.

On a complete network, the initial intervention is purely random for opinion-agnostic algorithms: as every agent has the same options available it it, every extremist is essentially indistinguishable from one another. After the first intervention, agents are then able to be differentiated due to the more complex network structure, but this structure is entirely reliant on the first random decision. This positive feedback loop then compounds itself. By moving to the Erdős–Rényi random network, we are able to explore a still simple graph that is less homogeneous than a complete network.

Test Case 2 *Targeted interventions - both opinion-aware and opinion-agnostic - will reduce polarisation on a random network and thus have a positive value for proportion saved, increasing as more edges are removed.*

Test Case 3 *Opinion-aware algorithms will have superior efficacy to opinion-agnostic algorithms.*

4.4 Metrics

In the following experiments we make use of two metrics. The first is a previously used metric introduced by Deffuant *et al.* and indicates the level and type of extremism within

an opinion dynamics model. The second is a new metric we propose for measuring the effectiveness of an intervention strategy.

4.4.1 Deffuant's y-metric

When generating heatmaps, we use the y-metric proposed by Deffuant *et al.* (Deffuant, Amblard, et al., 2002). We find p_+ and p_-, the proportion of agents whose final opinion was above 0.9 or below -0.9, respectively. The y-metric is then calculated using equation 4.1.

$$y = p_+^2 + p_-^2 \tag{4.1}$$

This results in $y = 1$ for single-extreme polarisation states, where either $p_+ = 1$ or $p_- = 1$, $y = 0.5$ for bimodal polarisation where $p_+ = p_- = 0.5$, and $y = 0$ for central convergence where $p_+ = p_- = 0$. Intermediate values indicate situations where full convergence was not achieved, and a number of agents became extreme but not all.

4.4.2 Proposal: Proportion Saved Metric

In order to measure the effect our intervention algorithms have on the network, we initially considered simply using the proportion of agents that did not become extreme by the end of the simulation: that is, $1 - b$, where b is the number of extremists. However, this metric would result in apparently high levels of success for every algorithm even when that algorithm was applied zero times. This is due to the metric counting those who never became extreme even without the algorithm as being "saved". As we see in figure 4.7, there are large portions of the heatmap in which some agents do not become extreme even without any intervention. It would be incorrect to apportion some of this prevention to our algorithms.

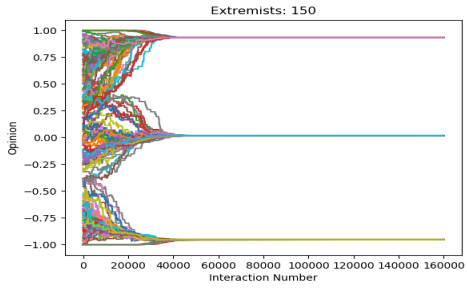
For each generated network, we instead calculate a, the number of initially non-extreme agents whose opinion at the end of the simulation has shifted to either above 0.9 or below -0.9 when we perform no interventions, as per the recommendations in Deffuant's work (Deffuant, Weisbuch, et al., 2013). We then reset the network to its initial state, perform the selected intervention algorithm, and then find b, again the number of newly-extreme agents. We are then able to find $1 - b/a$, the proportion of agents that would have become extreme had we done nothing, but did not become extreme as a result of our interventions. A value of 0 thus indicates that our interventions had no effect, values below 0 indicate that our interventions were worse than doing nothing, and a value of 1 indicating that every vulnerable agent was protected by our intervention. We report the mean and 95% confidence intervals of the 128 repetitions of the experiment.

In contrast to measuring the raw number of agents or proportion of agents that did not become extreme, this metric allows us to accurately measure the level of improvement offered by our intervention strategy over a baseline of doing nothing. It also standardises our results between different areas in the heatmap where different numbers of agents are exposed to extreme influence.

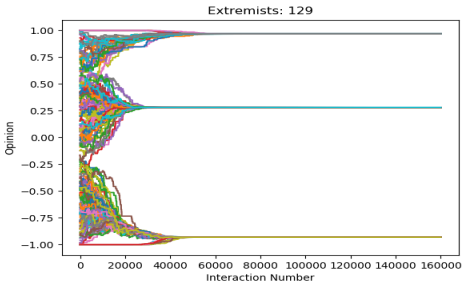
4.5 Complete Network Results

Figures 4.1 through 4.6 show typical results of applying the degree intervention algorithm 0, 1000, 2000, and 3000 times to several complete networks with different uncertainties and initial extremist populations. The opinion of each agent is shown from the beginning of the simulation through to the end, and illustrates that the interventions have very little effect. This is the case for every intervention algorithm attempted.

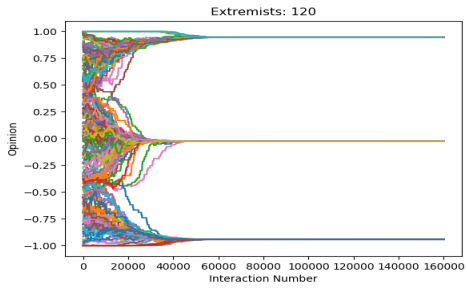
Hypothesis 1 proves to be correct. Figures 4.1 through 4.6 show representative samples from our experiments, and there is no significant difference between removing



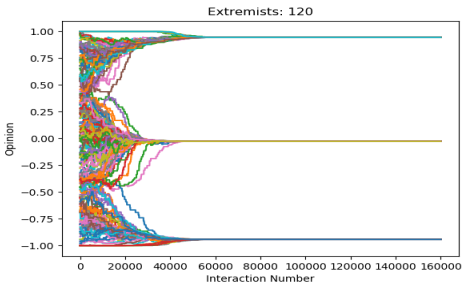
(a) 0 edges removed



(b) 1000 edges removed (approx. 2.5%)

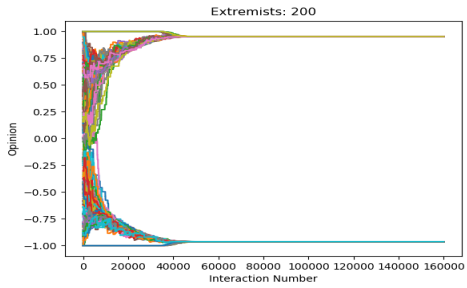


(c) 2000 edges removed (approx. 5.0%)

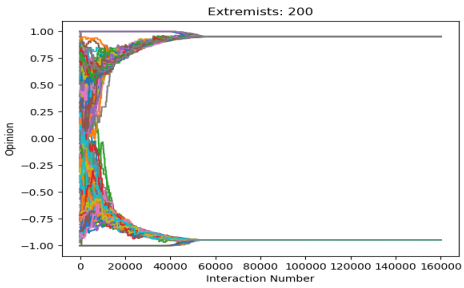


(d) 3000 edges removed (approx. 7.5%)

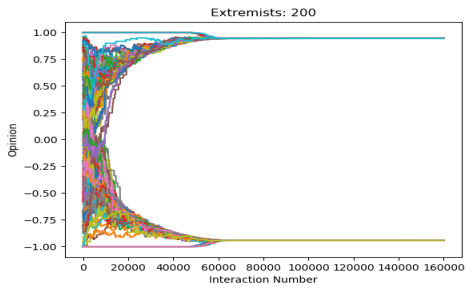
Figure 4.1: The evolution of opinion over time in a complete network is virtually unaffected by light-touch targeted removal of a given number of edges. This figure shows 0.5 uncertainty, 10% extremists.



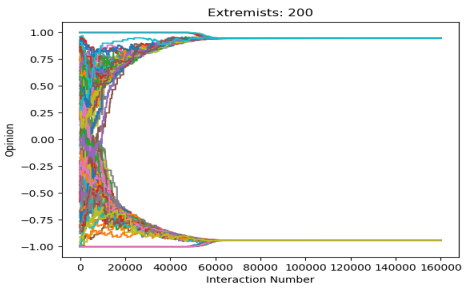
(a) 0 edges removed



(b) 1000 edges removed (approx. 2.5%)

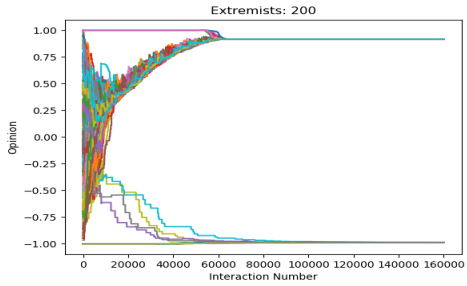


(c) 2000 edges removed (approx. 5.0%)

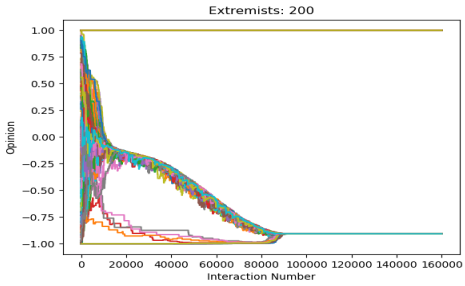


(d) 3000 edges removed (approx. 7.5%)

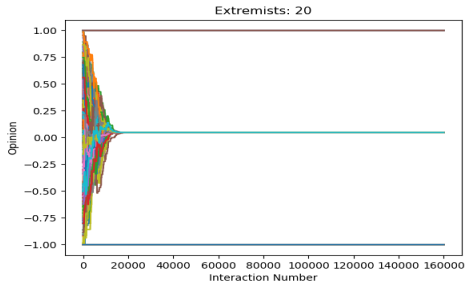
Figure 4.2: 1.0 uncertainty, 10% extremists.



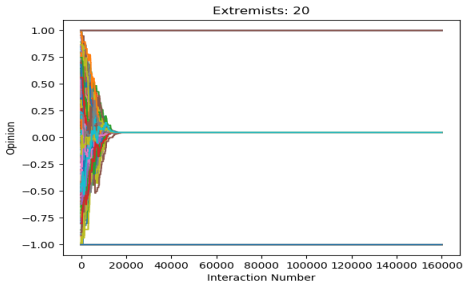
(a) 0 edges removed



(b) 1000 edges removed (approx. 2.5%)

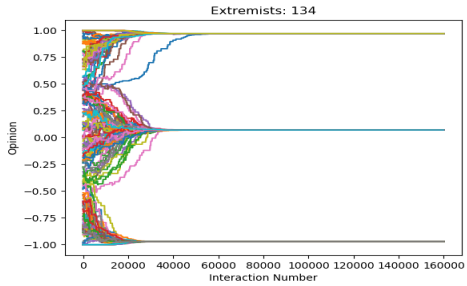


(c) 2000 edges removed (approx. 5.0%)

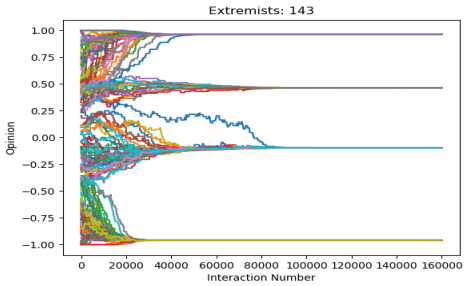


(d) 3000 edges removed (approx. 7.5%)

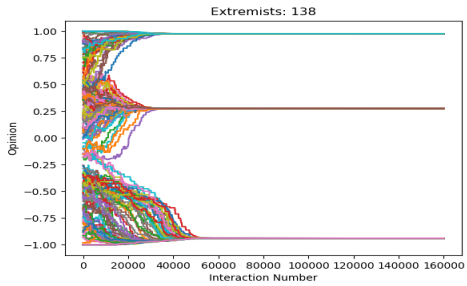
Figure 4.3: 1.5 uncertainty, 10% extremists.



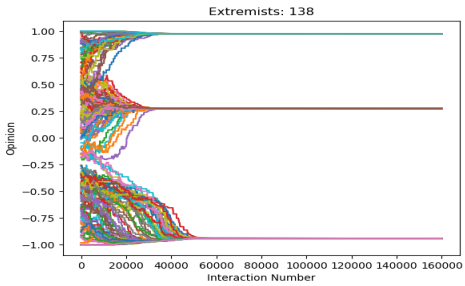
(a) 0 edges removed



(b) 1000 edges removed (approx. 2.5%)

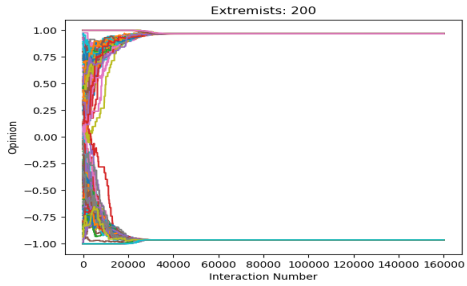


(c) 2000 edges removed (approx. 5.0%)

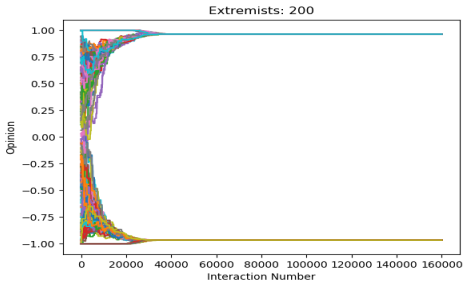


(d) 3000 edges removed (approx. 7.5%)

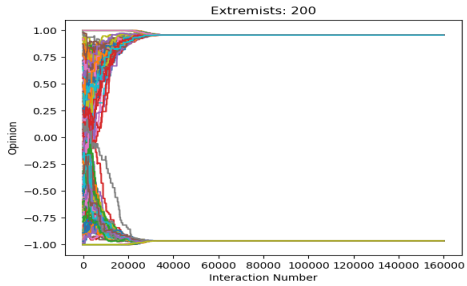
Figure 4.4: 0.5 uncertainty, 20% extremists.



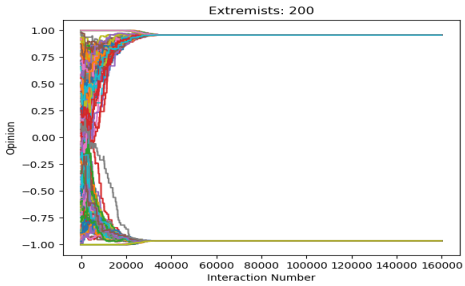
(a) 0 edges removed



(b) 1000 edges removed (approx. 2.5%)

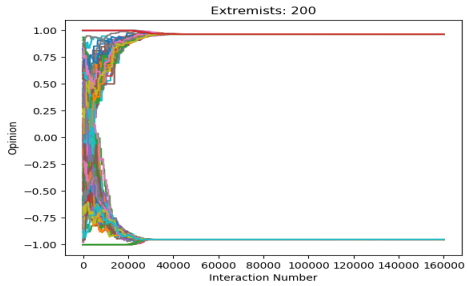


(c) 2000 edges removed (approx. 5.0%)

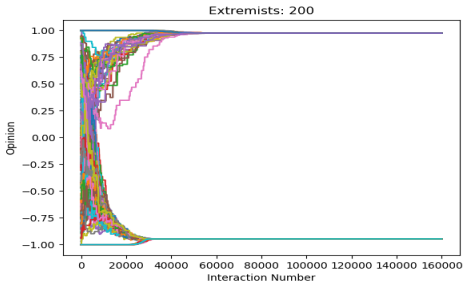


(d) 3000 edges removed (approx. 7.5%)

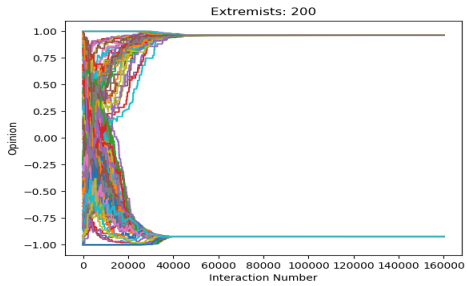
Figure 4.5: 1.0 uncertainty, 20% extremists.



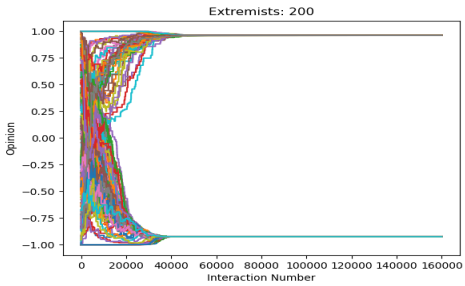
(a) 0 edges removed



(b) 1000 edges removed (approx. 2.5%)

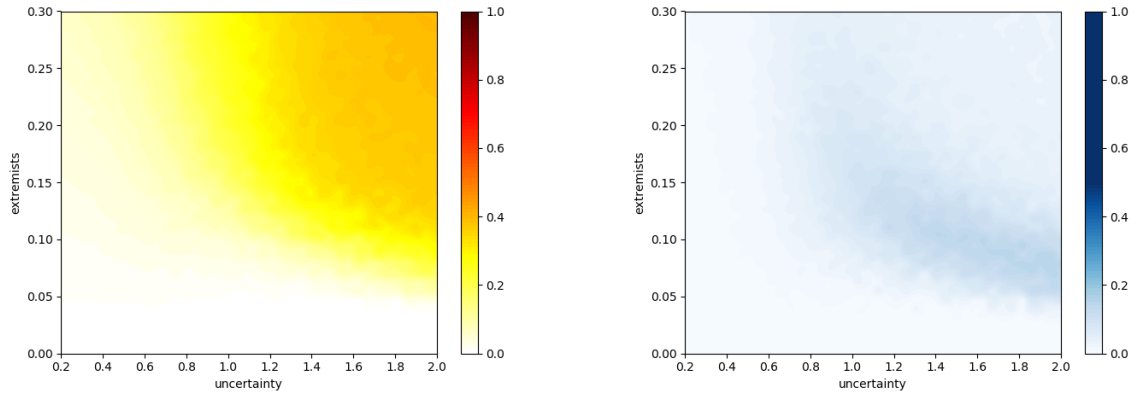


(c) 2000 edges removed (approx. 5.0%)



(d) 3000 edges removed (approx. 7.5%)

Figure 4.6: 1.5 uncertainty, 20% extremists.



(a) Mean **(b)** Standard Deviation

Figure 4.7: The y-metric for the relative agreement model on a random network with each agent having an average degree of 4.

zero and removing 7.5% of the edges in the network. It is clear that upon a complete network a light-touch intervention algorithm cannot meet with much success. Removing a relatively low number of edges still leaves extremists able to directly influence a substantial portion of the network. However, it is rare within a social network for every agent to be able to interact with every other agent. We thus turn our attentions to other network structures, such as the Erdős–Rényi $G(n, p)$ random network.

4.6 Random Network Results

Initial experimentation indicated that we would have to revise our method somewhat. The execution time for all but the algorithms of complexity $O(n)$ was considerable even on a relatively small network of 200 agents, and therefore would be impractical for larger networks. We therefore reduced our experimentation to use a subset of four algorithms: random, degree, and our proposed Vulnerability and Influence Targeting algorithms. This gives us a control case, a representative of opinion-agnostic algorithms, and a test of our contributions.

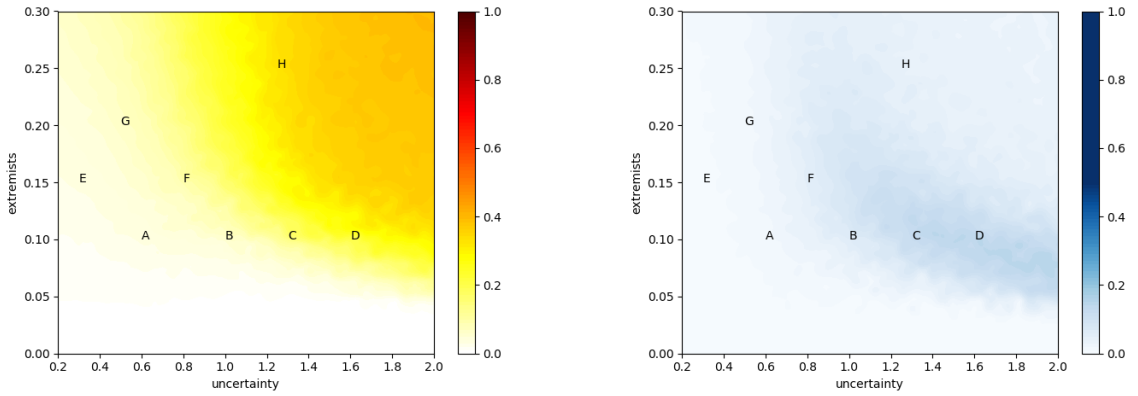
Figure 4.7(a) shows the mean y-metric for an ER $G(n, p)$ random graph with $n = 200$

and $p = 0.01$. This results in an average of 398 edges within the network, and so an average degree of approximately 4. We select 200 agents again so as to change only one thing at once. In changing the network generation algorithm, we do not also wish to change the size of the network at the same time. It is shown by Amblard *et al.* (Amblard and Deffuant, 2004) that sufficiently highly-connected networks show an almost inevitable drift to extreme polarisation. As we intend to investigate cases where extremism is not a foregone conclusion, we choose a relatively low level of connectivity to explore. As we intend to explore the Barabási–Albert scale-free network generator in chapter 6, we additionally ensured that p was such that the average number of edges was approximately a multiple of the number of nodes. This allowed us to generate scale-free networks with approximately the same number of edges. While the ER network does not accurately represent online social media, we use it as a stepping stone and a chance to investigate another well-known network structure.

4.6.1 Point Selection

It would be impractical to test our hypothesised intervention algorithms on every point within the heatmap, and so a number of indicative points were chosen upon which to test. These points were chosen to cover a wide range of different means and standard deviations. In addition, certain regions were identified within the heatmaps of mean and standard deviation, and points chosen both around the centre and the edges of these identified regions.

The points chosen for experimentation are shown in figure 4.8. We chose to focus our experimentation on areas outside the large orange component of the heatmap due to the inevitability of bi-modal convergence within that area. Regions where polarisation is less guaranteed are more likely to be susceptible to intervention. We consider four primary areas: the large orange section in the top-right, characterised by high y-metric



(a) Mean **(b)** Standard Deviation

Figure 4.8: The points on the heatmap selected for in-depth experimentation.

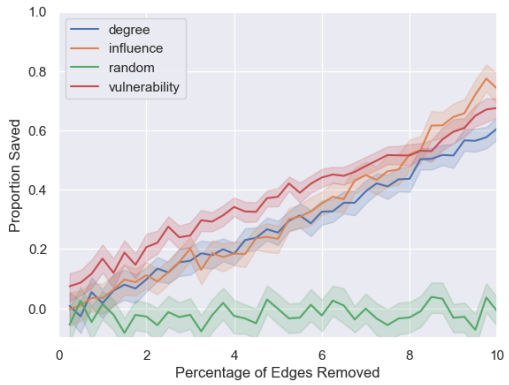
and low standard deviation; the yellow band surrounding it with high standard deviation and moderate y-metric; the pale yellow region to the left with low standard deviation and low y-metric; and finally the white band across the bottom, with low standard deviation and close to zero y-metric. Our points are selected to give us a good representation of the two yellow areas, where extremism can be observed (unlike the white area) without it becoming inevitable (unlike the orange area).

Points A $(0.6, 0.1)$ and G $(0.5, 0.2)$ were selected as they are representative of a region with a low y-metric that is quite far from a region of higher standard deviation. In this region we expect mostly central convergence, but with the occasional outlier. These outliers would ideally be caught and neutralised by our intervention algorithms.

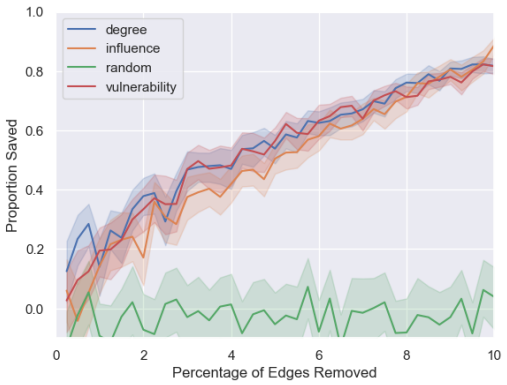
Points B $(1.0, 0.1)$ and F $(0.8, 0.15)$ have a higher number of extremists and a higher uncertainty. These approach the border of the medium y-metric region, and of the high standard deviation region, giving us an idea of what to expect as the simulation becomes less predictable.

Point C $(1.3, 0.1)$ continues on this progression, this time deep within the yellow band of medium y-metric and high standard deviation. At this point the simulation is highly unpredictable, with a standard deviation of above 0.18.

Point D $(1.6, 0.1)$ is similar to point C, though at this point we approach the border



(a) 10% extremists, 0.6 uncertainty



(b) 10% extremists, 1.0 uncertainty

(c) 10% extremists, 1.3 uncertainty

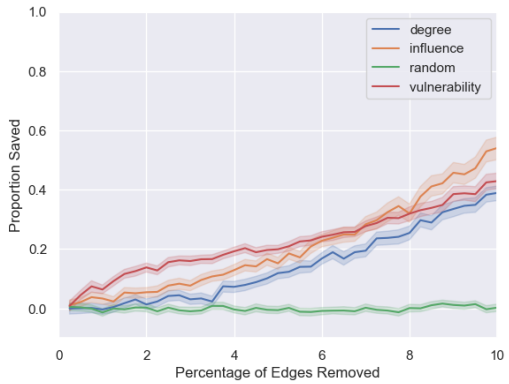
(d) 10% extremists, 1.6 uncertainty

Figure 4.9: Results for applying intervention algorithms to a random network show a reduction in network polarisation.

of the orange region, to investigate what happens on the brink of extremism becoming almost inevitably bi-modally polarised.

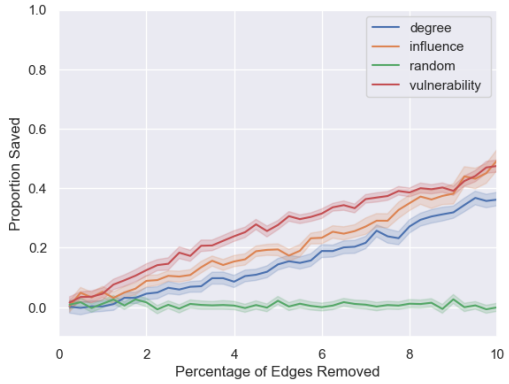
Point E $(0.3, 0.15)$ is well within the pale yellow area, exhibiting low y-metric as well as low extremism. In this area extremism is still reliably present though to a small degree. Our intervention algorithms should be able to reduce this extremism still further.

Point H $(1.25, 0.25)$ is a region with very high uncertainty and extremism. In this scenario a large number of edges must be severed before we see any real effect, and even at 10% of the edges removed, we only see around a 40% reduction in extremism.

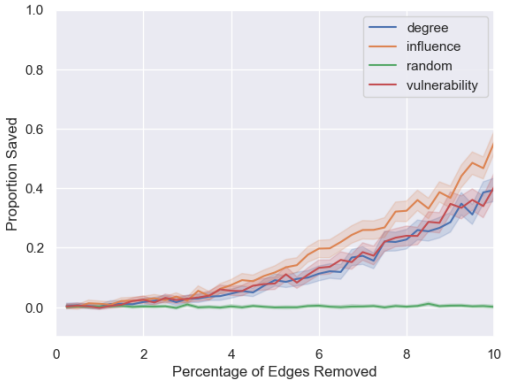


(e) 15% extremists, 0.3 uncertainty

(f) 15% extremists, 0.8 uncertainty



(g) 20% extremists, 0.5 uncertainty



(h) 25% extremists, 1.25 uncertainty

Figure 4.9: Results for applying intervention algorithms to a random network show a reduction in network polarisation.

4.6.2 Intervention Results

The results in figure 4.9 show a promising improvement over our chosen baseline, random edge removal, for both opinion-agnostic and opinion-aware algorithms.

Increasing the extremist proportion from 5% to 10% extremists in figures 4.9a through d, the results become more reliable. While at higher uncertainties in figures 4.9c and d all algorithms perform approximately the same, figure 4.9b shows a slight preference to the Vulnerability Targeting algorithm, and figure 4.9a a noticeable advantage to using opinion-aware algorithms. This advantage manifests strongly at low removal percentages for Vulnerability Targeting, while Influence Targeting performs particularly well after removing around 8% of the edges.

In figure 4.9g the difference becomes even more pronounced: after removing 5% of the edges, opinion-agnostic targeting saves roughly 10% of the vulnerable agents, whereas Vulnerability Targeting saves around 30% after removing the same number of edges.

Figure 4.9h shows very high proportions of extremists as well as initial uncertainty, and in this situation we only see a benefit over degree centrality using Influence Targeting. Even then, the improvement is marginal in such a hostile environment. Generally speaking, the Vulnerability Targeting algorithm has higher efficacy than Influence Targeting as well as the opinion-agnostic algorithms, though it is important to note that Influence Targeting often begins to outperform Vulnerability Targeting at higher intervention counts.

In figures A.1 and A.2 we show the proportion saved as well as raw number of extremists at regularly spaced points throughout the heatmap, to show how the simulation changes as we move through the parameter space.

4.7 Discussion

Hypothesis 2 is shown to be correct in all of the points selected. Even when uncertainty and initial extremist proportion is high, targeted intervention always outperforms random interventions. This has already been shown for other contagion and infection models (M. E. J. Newman, 2003), and this result is a strong indicator that mechanisms and systems for preventing infection in other domains can be successfully transferred to opinion dynamics models. Many of the points also demonstrate diminishing returns on efficacy, which supports our theory that a light-touch intervention may be able to substantially reduce network polarisation without too great a cost or impact on the network as a whole.

However, hypothesis 3 - that opinion-aware algorithms will outperform opinion-agnostic - only holds in limited circumstances. In those points where the uncertainty of agents is below 1, opinion-aware algorithms do indeed outperform opinion-agnostic algorithms. This is by a small margin in figures 4.9a and f, but by a substantial amount in figures 4.9e and g where the uncertainty is even lower.

In these scenarios with low uncertainty, only a subset of non-extreme agents are able to be influenced by extremists. We refer to this subset as vulnerables. Preventing extremist contact with vulnerable agents is key to containing the spread of polarisation almost before it starts, as it is these vulnerable agents that are able to influence non-vulnerable agents that the extremists cannot directly influence. With high levels of uncertainty the size of this subset of agents grows, and with uncertainty greater than 1 every agent is vulnerable as they can either by affected by one extremist or the other regardless of opinion. For high uncertainty then, opinion-aware algorithms offer no advantage. However, where uncertainty is low we are able to intervene only on vulnerable agents and thus not "waste" interventions by protecting those who are not vulnerable.

4.8 Conclusion

In this chapter, we demonstrated techniques that aim to minimise the spread of polarisation and extremism over a random network. This gave us an initial viewpoint into how these algorithms may behave on more realistic networks and allowed us to demonstrate the efficacy of these intervention methods.

We learned that when targeting based on opinion, an advantage over simple centrality-based targeting can be seen if and only if the uncertainty of the agents is below 1. This suggests that when the population is sufficiently vulnerable then it is best to target based on number of agents protected, whereas for more resilient communities targeting based on individual vulnerability may be best.

In the next chapter, we see if assigning different edges variable costs has an effect on the efficiency of intervention methods.

Chapter 5

Effects of Variable Edge Costs

So far in our experiments we have considered each edge to have the same cost to remove, whereas in real life this would not be the case. Removing a link between two casual acquaintances would be far less significant and have a smaller chance of being detected than removing a link between two relatives, for example. Intervening on the accounts of popular, famous, or politically sensitive people or organisations additionally carries costs above and beyond the costs of acting upon an average user.

It has already been seen that two of the largest online social networks consider certain actors or organisations exempt from rules due to political status and are unwilling to enforce rules equally to those agents (BBC News, 2019b; Twitter, 2019). In order to loosely model this disparity, we investigate the effects of assigning a cost to each edge and then instructing our algorithm to minimise cost, rather than minimising number of interventions.

In addition to the risk of political or social reprisal for action, social media organisations face a conflict of interest when it comes to reducing or preventing polarisation. Scandal, conflict, and outrage are good drivers for user engagement, which in turn directly influences advertising revenue. For instance, Twitter was able to profit from targeted advertising towards neo-Nazis until this was discovered and exposed by the BBC BBC

News, 2020.

It could be possible for several lower weighted edges to have a greater effect than a single highly weighted edge, yet have a lower overall chance of detection, and thus a lower cost. Our preference for a minimally costed intervention would then choose this set of several interventions, while our previous preference for a minimal number of edges severed would choose the single highly weighted edge. This initial exploration does not attempt complicated heuristics or solving schemes for optimising the set of choices, leaving that for future work. Instead we aim to see what effect a simple reward/cost judgement has using two example cost determination schemes, or pricing functions.

It is important to note that by cost we do not refer to a financial cost directly, but rather an abstract term for the political, business, and reputational consequences of performing an intervention. Intervening in the discussions between agents can be perceived as censorship or an assault on freedom of speech, and erroneous or not, could result in a loss of income or reputation for the organisation operating the intervention outcome. There is also the risk that a heavy-handed or over-zealous intervention can result in persons or groups of interest leaving the network altogether for a platform where they are less easily observed by law enforcement or intelligence.

5.1 Proposal: Edge Pricing Functions

Two algorithms were created to determine costs, based upon simple and local metrics for each agent, named *degree*, and *paths*. They do not attempt to objectively determine the importance of the relationship between two agents, but rather offer an approximation of how costly a given intervention would be, from 0 indicating no cost, to a maximum of 1. With a suitable cost and reward function in place, we could then be able to explore the resultant optimisation problem. In this initial exploration however, we use equation 5.1 to calculate a value for each edge, and sever the edge with the highest value.

$$\text{value} = \frac{\text{reward}}{\text{cost}} \tag{5.1}$$

Degree Degree pricing prices edges according to the degree of that edge's two endpoints, which is then normalised such that an edge between the highest degree agent and itself would be priced at 1. This is shown in equation 5.2, with $|i|$ and $|j|$ representing the number of neighbours of i and j, and $\max(N)$ representing the number of neighbours of the highest degree agent.

This algorithm places a high cost on edges that are likely to be observed, and thus their removal keenly felt. Users with many edges are assumed to be personally invested in the network, and possibly have analytical tools or social media teams watching them for any unusual activity. This makes removing edges from them a high-risk activity. Popular users are also assumed to be of relatively high importance to their followers, and so an agent finding themselves unable to view messages from a popular and well-known account would notice our intervention.

$$\text{cost} = \frac{|i| + |j|}{2\max(N)} \tag{5.2}$$

Paths The paths function prices edges between a given non-extreme agent and an extreme agent according to the proportion of the non-extreme agent's neighbours that were also neighbours of the extreme agent. In the case that two agents have no mutual neighbours, a cost of 0.01 is assigned to prevent division by zero. This is shown in listing 5.1.

This operates on the assumption that an edge disappearing from a previously tight-knit community of many mutual friends is a more serious impact than an edge disappearing from a sparse community. Families, cliques, and close-knit friendship groups also typical maintain offline links as well, and so removing these online links draws a great deal

Listing 5.1: Pseudocode for the Paths cost function

```
1 ni = neighbors(i) # non-extreme agent
2 nj = neighbors(j) # extreme agent
3
4 cost = sum(x in nj for x in ni)/len(ni)
5 cost = max(cost, 0.01)
```

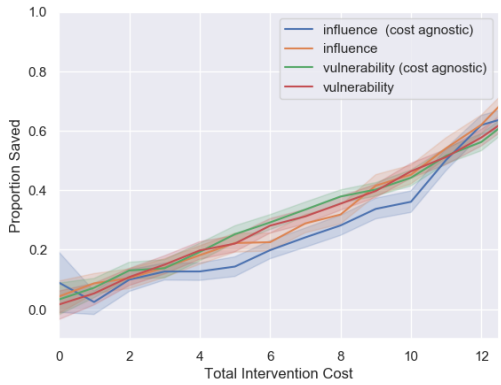
of suspicion. It may also be far less effective than the models demonstrate and thus cast doubt on any findings, due to the ability of those groups to try to re-establish communication through mechanisms outside of our algorithm's control.

It is important to note that costs are not comparable between these functions: a cost of 1 using the paths method could involve severing many more edges than the same cost using the degree method.

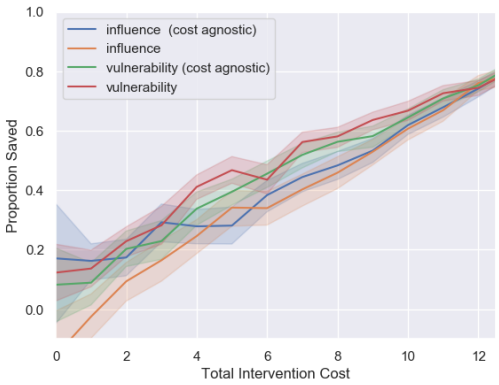
5.2 Experimental Setup

In setting up these experiments, we followed the same process as in section 4.2 with a small number of changes. Our intervention algorithms were modified to calculate costs according to our two defined edge pricing functions, and instead of choosing the edge with the highest weight, to choose the edge with the highest reward to cost ratio. We compare these results with those that do not take cost into account, which we call cost-agnostic. We continue to use the random network generator for these preliminary results.

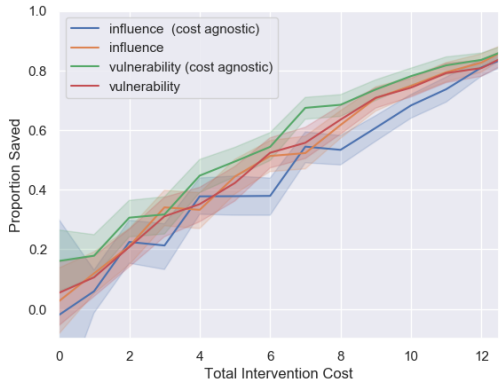
Both our targeting and our pricing algorithms are designed to have low computational complexity. Due to this there is substantial overlap in the information used to estimate value as well as that used to calculate cost. This renders some combinations invalid, such as Degree targeting with Degree costing. In this case the adjusted reward/cost for every edge would be 1. For this reason and to aid with legibility, we have omitted degree and



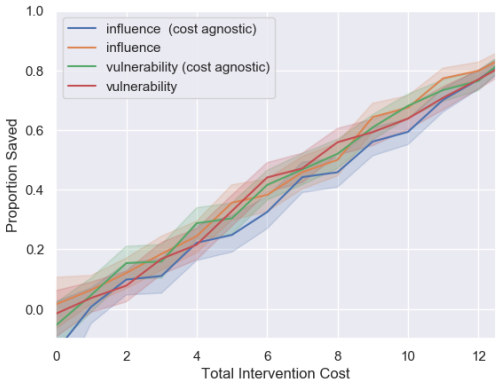
(a) 10% extremists, 0.6 uncertainty



(b) 10% extremists, 1.0 uncertainty



(c) 10% extremists, 1.3 uncertainty



(d) 10% extremists, 1.6 uncertainty

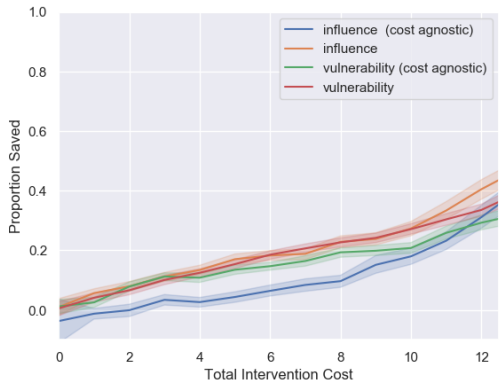
Figure 5.1: Results for applying our intervention algorithms to a random network after accounting for costs based on the degrees of the agents involved.

random targeting from the graphs below, and show only our algorithms in competition with each other and their cost-agnostic versions.

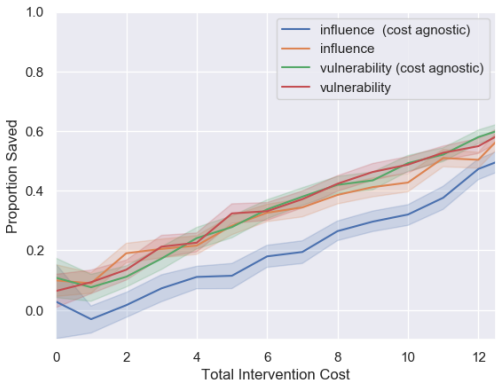
Test Case 4 *Cost-aware algorithms will be more cost efficient than cost-agnostic algorithms, as they will avoid high-reward/high-cost edges in favour of medium-reward/low-cost edges.*

5.3 Results and Discussion

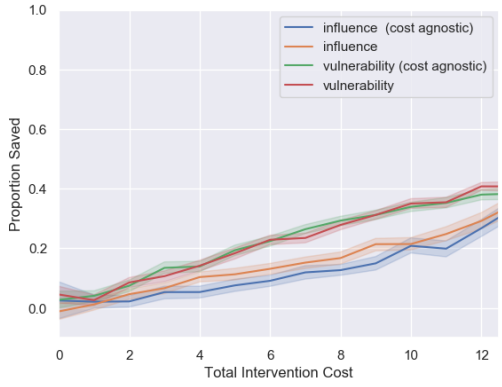
Our results, shown in figures 5.1 and 5.2 display a very small performance advantage to cost-agnostic algorithms with low proportions of edges removed, though this advantage



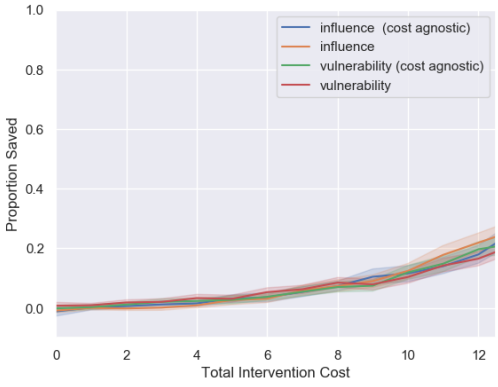
(e) 15% extremists, 0.3 uncertainty



(f) 15% extremists, 0.8 uncertainty

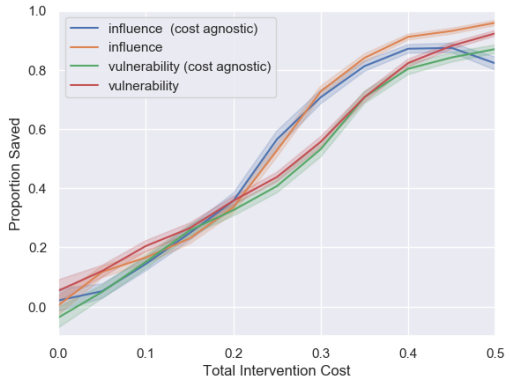


(g) 20% extremists, 0.5 uncertainty

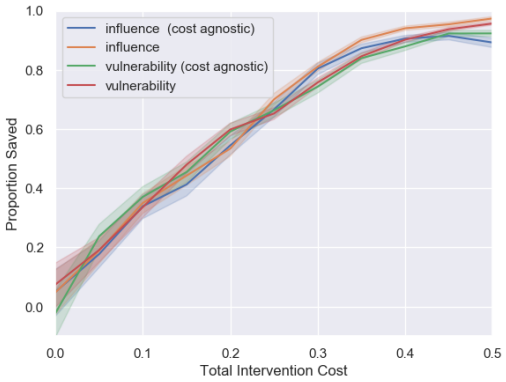


(h) 25% extremists, 1.25 uncertainty

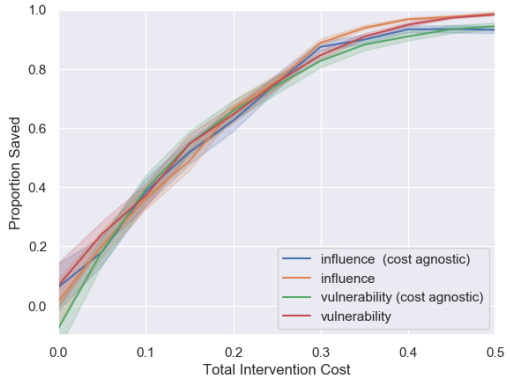
Figure 5.1: Results for applying our intervention algorithms to a random network after accounting for costs based on the degrees of the agents involved.



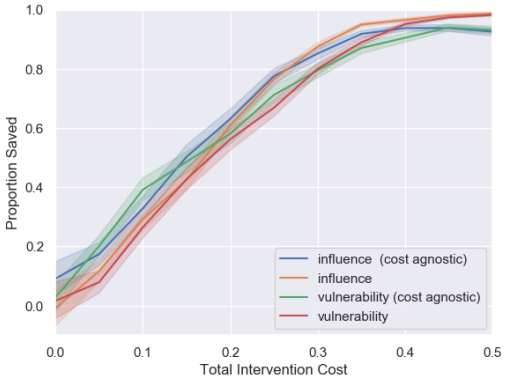
(a) 10% extremists, 0.6 uncertainty



(b) 10% extremists, 1.0 uncertainty

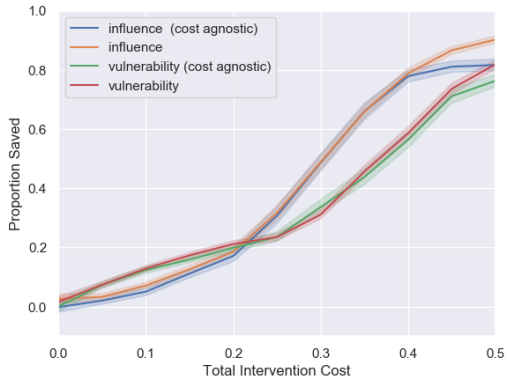


(c) 10% extremists, 1.3 uncertainty

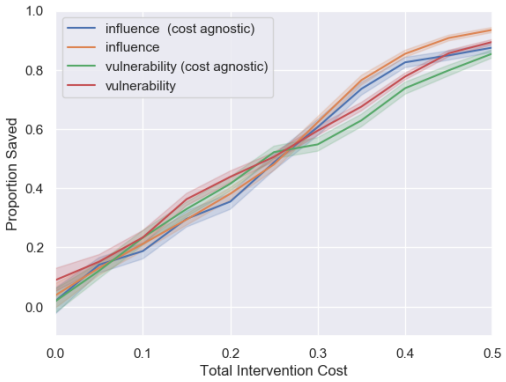


(d) 10% extremists, 1.6 uncertainty

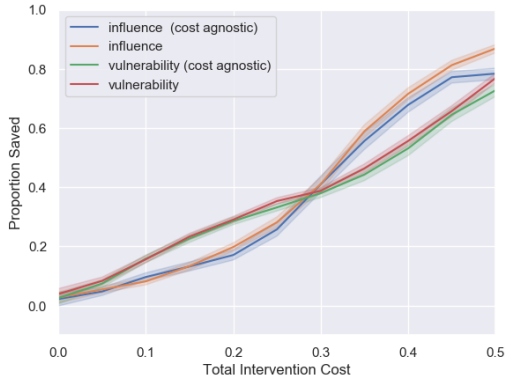
Figure 5.2: Results for applying our intervention algorithms to a random network after accounting for costs based on the mutual friends of the agents involved.



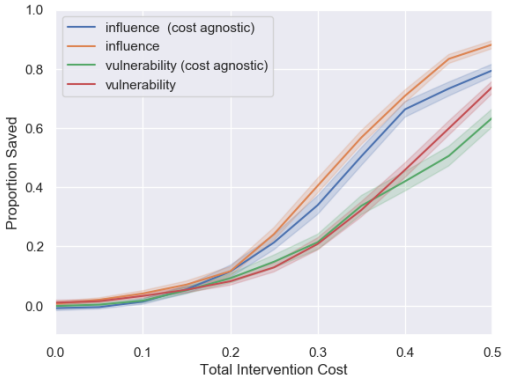
(e) 15% extremists, 0.3 uncertainty



(f) 15% extremists, 0.8 uncertainty



(g) 20% extremists, 0.5 uncertainty



(h) 25% extremists, 1.25 uncertainty

Figure 5.2: Results for applying our intervention algorithms to a random network after accounting for costs based on the mutual friends of the agents involved.

quickly drops off as more edges are removed. For sufficiently large proportions of edges removed, around 10%, the set of edges selected for removal remains the same whether cost is taken into account or not: applying these costs to edges appears to only change the order in which interventions are made, not the interventions themselves. Cost-agnostic algorithms are able to perform high-cost, high-reward interventions immediately, while cost-aware algorithms are forced to choose those with higher reward ratios but not necessarily higher gross returns.

We note that in path-based pricing structures, efficacy of cost-agnostic algorithms drops at higher costs compared to cost-aware algorithms, which always perform better with higher costs. The cost-aware algorithms show a strong preference for totally disconnecting agents from the network where possible: when an extremist has only one edge remaining, that edge is deemed to have a very low cost due to it being impossible for any common neighbours to exist. This points to an advantage to be gained by carefully managing the connections of high-degree extremists due to the high cost of operating on them, while deleting extremist accounts with few followers due to the low cost of acting.

Finally, we note that within a random network agents are likely to have close to the mean number of edges - in this case, 4. In future work we could explore the differences between cost-agnostic and cost-aware algorithms on networks with different degree distributions, such as the scale-free networks described elsewhere within this thesis. Due to this not being the core focus of the work however, we elected to limit our research into costing methodologies to random networks at this point.

5.4 Comparing Costing Mechanisms

The next part of our analysis is to investigate to what extent the pricing algorithms agree upon ideal targets to select. It can be seen in figures 5.1 and 5.2 that performance is similar in many cases between cost-aware and cost-agnostic versions of the intervention

Listing 5.2: Code for the Influence Targeting algorithm

```
rem_a = //list of edges removed by algorithm a
rem_b = //list of edges removed by algorithm b
numb = len(rem_a)
overlap = sum(x in rem_b for x in rem_a)/numb
```

0.1 ex, 0.6 un	Vuln ag	Vuln deg	Vuln pat	Inf ag	Inf deg	Inf pat
Vuln agnostic	1	0.85	0.75	0	0	0
Vuln degree	0.85	1	0.9	0	0	0
Vuln paths	0.75	0.9	1	0	0	0
Int agnostic	0	0	0	1	0.8	0.45
Int degree	0	0	0	0.8	1	0.55
Int paths	0	0	0	0.45	0.55	1

Table 5.1: 10% extremists, 0.6 uncertainty

algorithms. To investigate this, for each point we investigated within the heatmap we created six identical graphs using the same random seed, and tested each of our intervention algorithms upon a graph using degree costing, path costing, and cost-agnostic interventions. We then inspected the first 20 removals by each algorithm, and calculated the overlap between these lists using listing 5.2. This code returns the proportion of edges removed by the first algorithm that were also removed by the second algorithm. 20 was chosen as this represents 5% of the total edges removed, which we consider to represent a light-touch intervention. This was then repeated 20 times, and the average overlap taken.

In tables 5.1 through 5.7, values of 1 indicate that the pair of algorithms are in total agreement on the first 20 edges to remove, and values of 0 indicate that there are no common targets between the two generated lists. Owing to page width constraints, Influence Targeting has been abbreviated to "Inf" and Vulnerability Targeting to "Vuln".

One of the most immediately striking features in each table is that there is virtually no agreement between Vulnerability and Influence Targeting on which agents and edges to target. Despite this, they have broadly the same success rate as one another. This

0.1 ex, 1.0 un	Vuln ag	Vuln deg	Vuln pat	Inf ag	Inf deg	Inf pat
Vuln agnostic	1	1	0.9	0	0	0
Vuln degree	1	1	0.9	0	0	0
Vuln paths	0.9	0.9	1	0	0	0
Int agnostic	0	0	0	1	0.85	0.6
Int degree	0	0	0	0.85	1	0.65
Int paths	0	0	0	0.6	0.65	1

Table 5.2: 10% extremists, 1.0 uncertainty

0.1 ex, 1.3 un	Vuln ag	Vuln deg	Vuln pat	Inf ag	Inf deg	Inf pat
Vuln agnostic	1	0.9	0.75	0	0	0
Vuln degree	0.9	1	0.85	0	0	0
Vuln paths	0.75	0.85	1	0	0	0
Int agnostic	0	0	0	1	0.7	0.65
Int degree	0	0	0	0.7	1	0.85
Int paths	0	0	0	0.65	0.85	1

Table 5.3: 10% extremists, 1.3 uncertainty

0.1 ex, 1.6 un	Vuln ag	Vuln deg	Vuln pat	Inf ag	Inf deg	Inf pat
Vuln agnostic	1	1	1	0.05	0.05	0
Vuln degree	1	1	1	0.05	0.05	0
Vuln paths	1	1	1	0.05	0.05	0
Int agnostic	0.05	0.05	0.05	1	0.9	0.45
Int degree	0.05	0.05	0.05	0.9	1	0.5
Int paths	0	0	0	0.45	0.5	1

Table 5.4: 10% extremists, 1.6 uncertainty

0.15 ex, 0.3 un	Vuln ag	Vuln deg	Vuln pat	Inf ag	Inf deg	Inf pat
Vuln agnostic	1	0.85	0.75	0	0	0
Vuln degree	0.85	1	0.75	0	0	0
Vuln paths	0.75	0.75	1	0	0	0
Int agnostic	0	0	0	1	1	0.5
Int degree	0	0	0	1	1	0.5
Int paths	0	0	0	0.5	0.5	1

Table 5.5: 15% extremists, 0.3 uncertainty

0.15 ex, 0.8 un	Vuln ag	Vuln deg	Vuln pat	Inf ag	Inf deg	Inf pat
Vuln agnostic	1	0.85	0.75	0	0	0
Vuln degree	0.85	1	0.9	0	0	0
Vuln paths	0.75	0.9	1	0	0	0
Int agnostic	0	0	0	1	0.85	0.4
Int degree	0	0	0	0.85	1	0.45
Int paths	0	0	0	0.4	0.45	1

Table 5.6: 15% extremists, 0.8 uncertainty

0.2 ex, 0.5 un	Vuln ag	Vuln deg	Vuln pat	Inf ag	Inf deg	Inf pat
Vuln agnostic	1	0.95	0.85	0	0	0
Vuln degree	0.95	1	0.9	0	0	0
Vuln paths	0.85	0.9	1	0	0	0
Int agnostic	0	0	0	1	0.8	0.4
Int degree	0	0	0	0.8	1	0.45
Int paths	0	0	0	0.4	0.45	1

Table 5.7: 20% extremists, 0.5 uncertainty

0.25 ex, 1.25 un	Vuln ag	Vuln deg	Vuln pat	Inf ag	Inf deg	Inf pat
Vuln agnostic	1	0.9	0.7	0	0	0
Vuln degree	0.9	1	0.75	0	0	0
Vuln paths	0.7	0.75	1	0	0	0
Int agnostic	0	0	0	1	0.65	0.45
Int degree	0	0	0	0.65	1	0.45
Int paths	0	0	0	0.45	0.45	1

Table 5.8: 25% extremists, 1.25 uncertainty

would suggest that both Vulnerability and Influence Targeting, despite their different intentions and modes of operation, both have their place in an intervention strategy. This reinforces our suggestion in section 4.1, that future developments should consider both probability and consequence of radicalisation when deciding how best to intervene.

Besides the inevitable case of each algorithm/cost pair having perfect agreement with itself, the next strongest agreement is between each of the different variants of the Vulnerability Targeting algorithm. At all points tested, cost-aware and cost-agnostic pricing schemes had at least a 75% agreement with one another, meaning that agnostic, degree, and paths-based edge selection produced largely the same list of edges to be removed. This indicates that under this intervention algorithm, certain interventions are worth performing no matter the cost. This explains the very close alignment in performance between cost-agnostic and cost-aware versions of Vulnerability Targeting: they are performing mostly the same actions. There is however some variation in the order in which these actions are taken.

There is a similarly strong agreement between degree-costed and cost-agnostic Influence Targeting, with at least 70% agreement between the two pricing strategies. Interestingly however, this did not hold for paths-based costing. While a few edges with extremely high scores appeared in both paths-costed and cost-agnostic intervention lists, the majority of the lists differed. The fact that despite the differences in these lists, the two strategies had almost equal performance suggests that once these high-value edges are severed, further interventions were less important in where they occurred, only that they did occur.

5.5 Discussion

These experiments show that while assigning varying costs to edges may lead to some benefits as seen in figures 5.1e, f, and g, in many cases these proposed measures do

not materially impact the success rate or cost-efficiency of intervention strategies. The strong degree of overlap shown in tables 5.1 through 5.8 demonstrate that in many cases the expected yield of removing a certain edge far outweighs the cost regardless of costing mechanism. This goes directly against hypothesis 4. We also note that the degree costing algorithms frequently prioritises nodes with very few neighbours, often leaf nodes. While on a random network all nodes will have approximately the same cost, on a scale free network there are orders of magnitude of difference between the highest and least connected agents. In future work we could investigate a costing mechanism that considers logarithms of the degree rather than the raw degrees of agents, thus reducing the effects of such a drastic difference in degree between popular and unpopular actors within the network.

Pricing based purely on network structural effects, at least in these cases, appears to have made little difference overall. In future work we would like to explore larger networks where there are more choices of edges to cut, and investigate whether the high level of overlap exists there. Implementing some costs independent of the network structure would also allow us to investigate a more realistic environment, and thus generate insights more directly applicable to the real world.

5.6 Conclusion

In this chapter we explored two simple edge costing procedures: weighting edges according to the degree of the endpoints of those edges, and by the proportion of mutual neighbours those endpoints had. We chose two simple metrics for this, based on either the degree of the involved agents or the proportion of neighbours they had in common. These stood in as proxies for two different factors in determining the importance of a link, the popularity of the agents involved and to what extent that link exists in a tightly-knit community. We then instructed our intervention algorithms to target those edges with the highest

reward to cost ratio. However, it is important to note that these pricing mechanisms only take into account network structural factors. In a real-world scenario there are likely to be additional factors independent of the network structure that have an effect on the cost - financial, temporal, or political - of performing a particular intervention.

We found that contrary to our hypothesis, there was very little improvement in cost effectiveness using either of these techniques compared to assuming all edges have a uniform cost. We then explored this further and found that in the majority of cases the interventions being performed were the same regardless of cost. Many edges had such a high reward that the cost was irrelevant. We also note some limitations of our approach that we would like to explore in future work, such as the relatively small network size and the homogeneity of the agents within our chosen network structure.

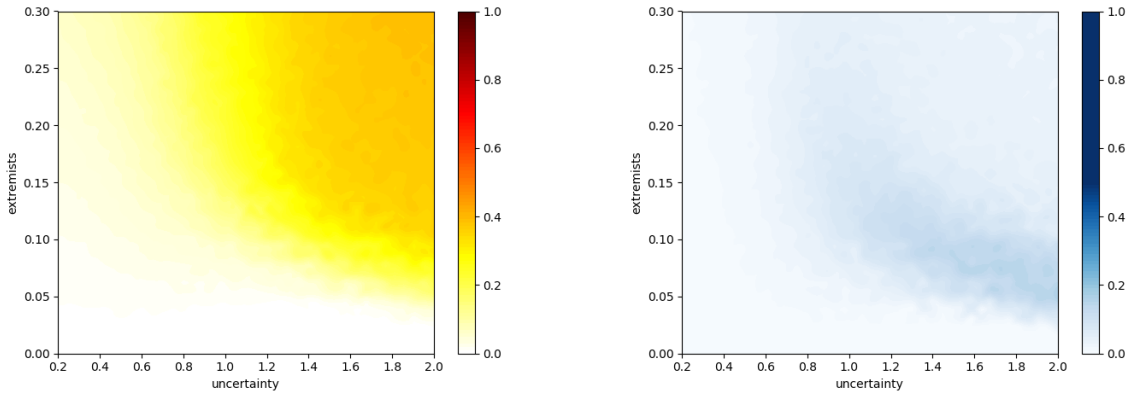
In the following chapter we increase the realism of the model through another angle. Where this chapter explored altering the intervention algorithms, Chapter 6 investigates the effects of changing network creation rules to mirror the structure of social networks such as Twitter and Facebook. Chapter 6.4 extends further on this theme and changes the communication and update rules of the model to more accurately capture the dynamics of public "one to many" messaging platforms such as Twitter. This marks a significant departure from the one to one communications assumed previously in this thesis.

Chapter 6

Applying Intervention Algorithms to Scale-Free Networks

Chapter 4 showed us that intervention methods can have a significant effect on the emergence of extremism and polarisation in social networks. However, it remains to be seen whether these algorithms function on networks designed to have similar properties to those found in the real world. It is possible that these successes are an artefact of the network structure and only arise under those specific conditions in much the same way as single extreme convergence only arises in highly connected networks (Amblard, Weisbuch, and Deffuant, 2003).

In this chapter, we apply the intervention algorithms to two types of scale-free networks. Networks are created using the Barabási–Albert preferential attachment model in each case, but the method in which they are seeded with opinions varies. In BA networks, agents are created and then immediately create a number of edges to other agents, preferring to select those agents that already have more edges connected to them. This forms a self-reinforcing loop akin to the phrase "the rich get richer", in which agents with many edges attached are more likely to gain still more edges. We select 25% of the agents with the highest degrees, which we refer to as *hubs*. We select the top 25%



(a) Mean **(b)** Standard Deviation

Figure 6.1: The y-metric for the relative agreement model on a scale-free network with extremist hubs.

to ensure that the entire tail of the degree distribution is caught, while leaving most of the body of the distribution outside our control. We then generate networks with the constraint that none of these hubs may be extremist, or the constraint that at least one of these hubs must be a positive extremist and at least one a negative extremist. This eliminates the unlikely possibility that a network is generated with a highly imbalanced level of influence between sides, which would skew the y-metric.

6.1 Extremist Hubs

As shown by figure 6.1, guaranteeing both positive and negative extremists a position within the top 25% most connected agents results in a very similar heatmap to that displayed in figure 4.7. This is due to the relatively high likelihood of both positive and negative extremists being allocated a hub by pure chance. Guaranteeing this outcome only changes it from being highly likely, to being certain. Below approximately 5% extremists, extremism is unable to get a foothold within the network except with very high uncertainties. As the proportion of extremists grows, so too does the y-metric and thus the proportion of agents converted to extremism. However, a degree of uncertainty

is required for extremism to fully propagate throughout the network.

With this opinion dynamics model, extremists must rely on being directly connected to a target agent, or else radicalising one who they are directly connected to and having them influence the target. There is no mechanism for directly interacting with an agent that you are not connected with. Increasing the number of extremists increases the probability of an extremist being directly connected to a vulnerable agent through increasing the number of edges within the network that connect to an extremist, while increasing the uncertainty increases the probability by increasing the number of vulnerable agents.

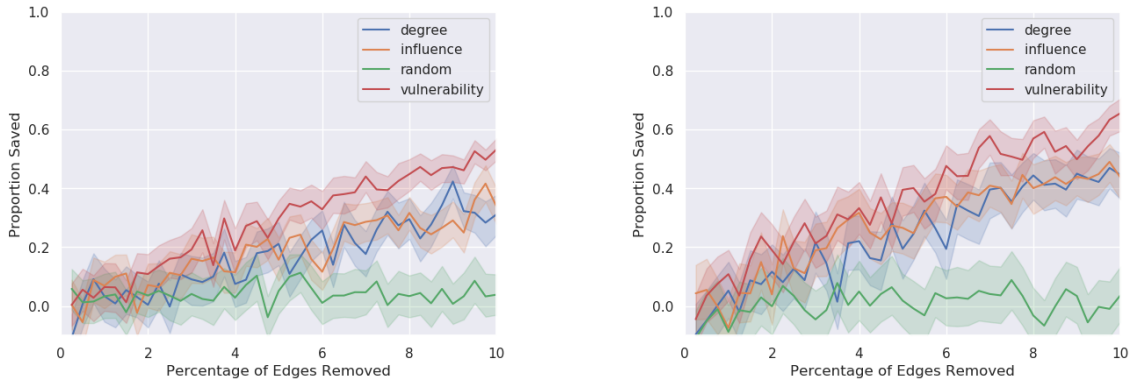
Due to the similarity of this heatmap to that in the previous chapter, we elected to use the same points for experimentation and analysis. This additionally allows us to compare the same points across multiple graph generation rules.

Test Case 5 *Similarly to random networks, opinion-aware algorithms will outperform opinion-agnostic algorithms when the uncertainty is less than 1.*

Results from experimentation on these scale-free graphs are shown in figure 6.2. Once again the Vulnerability Targeting algorithm displays superior efficacy to other opinion-aware and opinion-agnostic algorithms in figures 6.2a, e, f, and g.

The curves produced by plotting proportion saved against proportion of edges removed broadly fall into one of two types. The first can be seen in figures 6.2b, c, d, and to a lesser extent f. These show an initially fast yet decelerating growth, reminiscent of a logarithmic function.

The other form is seen in figures 6.2e, g, and to some degree a. These graphs show the inverse of the first form for opinion-agnostic and Influence Targeting, in that they perform poorly at low removal percentages but show accelerating growth. Vulnerability Targeting however maintains the initial strong performance at low proportions of edges removed. These points are all at areas within the heatmap with low standard deviations, indicating reliable and predictable extremist performances. In areas with a higher standard



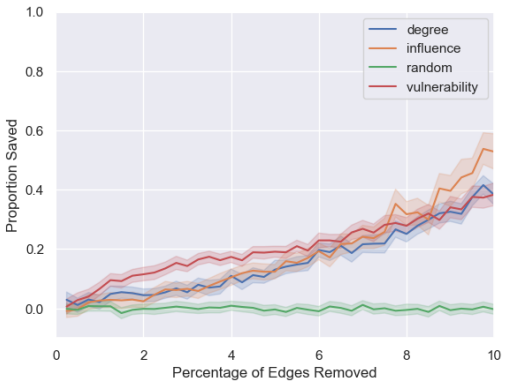
(a) 10% extremists, 0.6 uncertainty

(b) 10% extremists, 1.0 uncertainty

(c) 10% extremists, 1.3 uncertainty

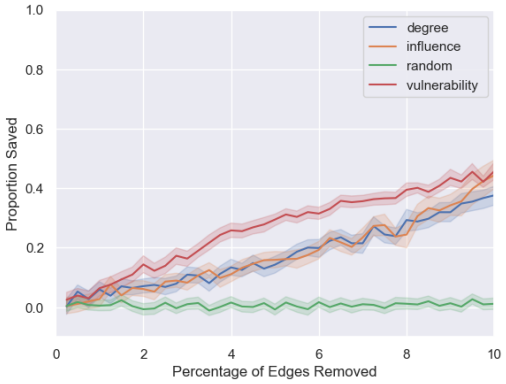
(d) 10% extremists, 1.6 uncertainty

Figure 6.2: Results for applying intervention algorithms to a scale-free network where each side has at least one highly-connected extremist agent.

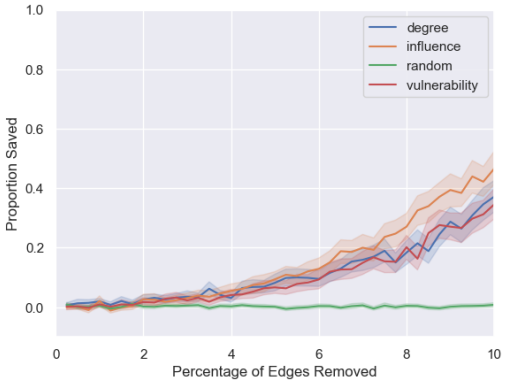


(e) 15% extremists, 0.3 uncertainty

(f) 15% extremists, 0.8 uncertainty

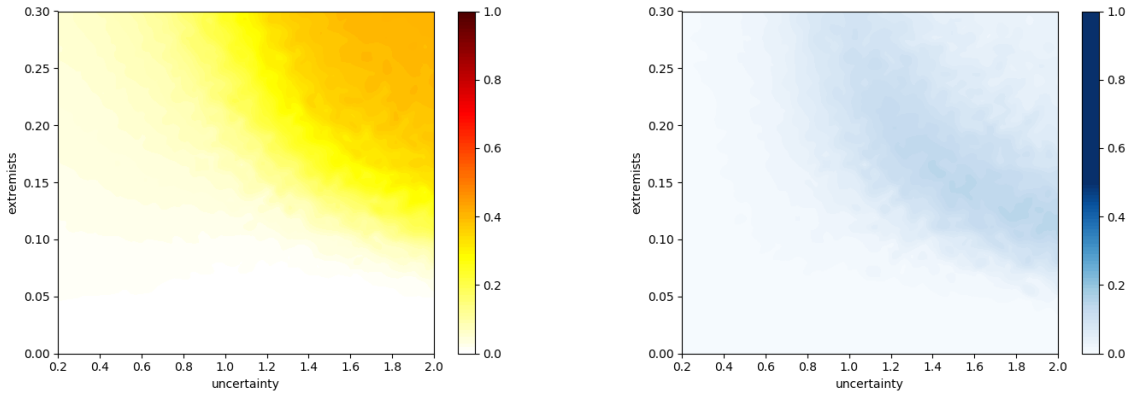


(g) 20% extremists, 0.5 uncertainty



(h) 25% extremists, 1.25 uncertainty

Figure 6.2: Results for applying intervention algorithms to a scale-free network where each side has at least one highly-connected extremist agent.



(a) Mean **(b)** Standard Deviation

Figure 6.3: The y-metric for the relative agreement model on a scale-free network with no extremist hubs.

deviation, the first form of curve is seen instead.

In keeping with figure 4.9h, figure 6.2h also shows only a slight improvement to Influence Targeting over degree-based targeting methods.

Of note is that the opinion-agnostic algorithms performed better on this scale-free network than they did on the random network. Rather than performing equally to opinion-aware as in figure 4.9b, figure 6.2b shows the opinion-agnostic methods perform fractionally better. A similar phenomenon is observed in figures 6.2f and g, where opinion-agnostic methods perform equally to Influence Targeting.

Figure A.5 shows trends as we move through parameter space. The proportion of extremists grows as we move upwards and to the right, representing increased initial extremist population and increased uncertainty. Our interventions have weakened efficiency at increasing uncertainty, but are still able to show a performance increase over opinion-agnostic methods.

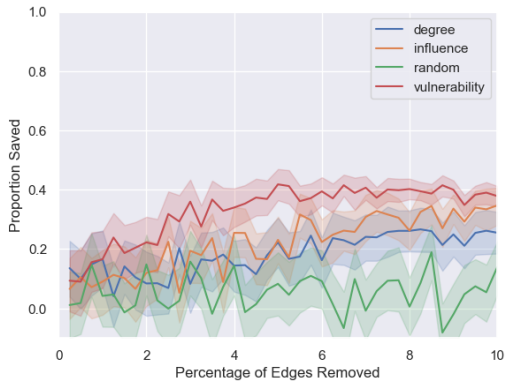
6.2 No Extremist Hubs

Our investigation for this section focuses on so-called "third spaces", communities in which the primary topic of discussion is centred on a mostly apolitical topic. These spaces are characterised by rational and civil debate between members from across the political spectrum (Wright, Graham, and Jackson, 2017). It thus follows that these spaces are desirable targets for both politically-motivated and purely recreational trolling.

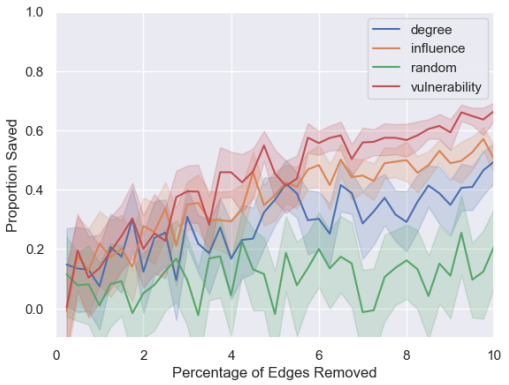
In this experiment, we consider these spaces to be established with no extremists initially, and for extremists to then be added to the system. This represents an established community becoming targeted by extremists from outside. These outside attackers thus do not have the advantage of having highly-connected agents initially within the network. We simulate this by enforcing that none of the top 25% most connected agents can be extremists at the beginning of the simulation. The choice of 25% is to ensure that nobody within the long tail of a scale-free network's degree distribution can be extreme. We estimate that the presence of even a single extreme hub will drastically alter the results, while incorrectly classifying an agent as a hub will not have such a drastic effect. In future work we would be interested in exploring different cut-off points or alternative methods of classification.

Test Case 6 *Opinion-aware algorithms will significantly outperform opinion-agnostic algorithms on third space-type networks.*

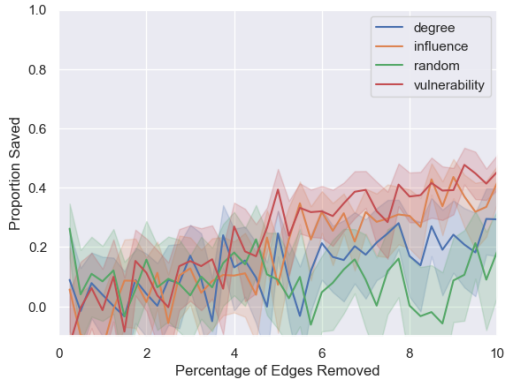
We hypothesise that opinion-aware techniques will exhibit greater prevention capabilities compared to agnostic techniques on this type of structure, and that the difference in performance will be greater than that seen on scale-free networks in which extremist accounts have already reached prominence, as in section 6.1. In these circumstances agents appear more similar to opinion-agnostic algorithms, as there are fewer outlying extremists with far more edges than others than would usually be found in the higher



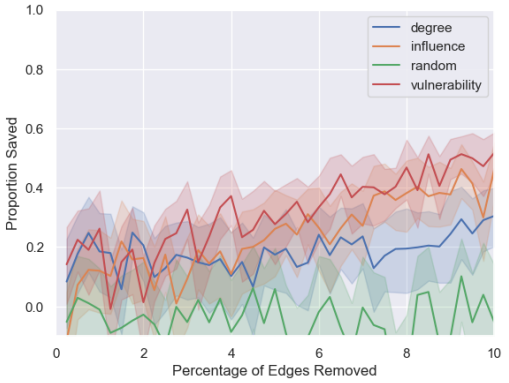
(a) 10% extremists, 0.6 uncertainty



(b) 10% extremists, 1.0 uncertainty



(c) 10% extremists, 1.3 uncertainty



(d) 10% extremists, 1.6 uncertainty

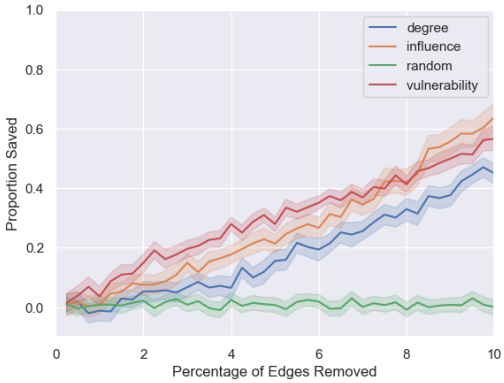
Figure 6.4: Results for applying intervention algorithms to a scale-free network where neither side has a highly connected agent.

end of the degree distribution of a scale-free network. This lends an advantage to opinion-aware algorithms which are able to differentiate between clusters of agents with a strong structural similarity.

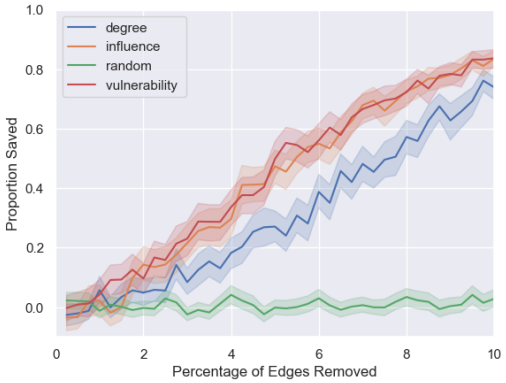
The most immediately obvious feature of the figures in figure 6.4 when compared to those in figure 6.2 is the considerably higher variance in all instances. Despite having the same number of repetitions of the experiment, the confidence intervals are significantly larger. This is due to the reach of each extremist faction being far more variable than when each side is guaranteed a hub. In situations where the uncertainty is very high such as 6.4c and d, the reach of extremists is far more important than the original alignment of non-extremists, as they are likely to be persuaded by any extremist they come into

(e) 15% extremists, 0.3 uncertainty

(f) 15% extremists, 0.8 uncertainty



(g) 20% extremists, 0.5 uncertainty



(h) 25% extremists, 1.25 uncertainty

Figure 6.4: Results for applying intervention algorithms to a scale-free network where neither side has a highly connected agent.

contact with. This allows the higher variance in reach to translate directly into a higher variance in conversion rate of neutrals to extremists.

This difference in reach and uncertainty also explains why opinion-aware algorithms do not substantially outperform opinion-agnostic algorithms in situations of high uncertainty. In these scenarios, all agents can be considered vulnerable and so targeting based on vulnerability is less effective. When almost all agents can be persuaded by extremists, it is more effective to target based solely on reach and audience size, and thus centrality.

It is apparent that opinion-agnostic algorithms perform substantially less well: in figure 6.4c, removing 5% of the edges resulted in saving 40% of the vulnerable agents, down from saving 60% in figure 6.2c. By contrast, opinion-aware algorithms equal or exceed their performance in scale-free networks with extremist hubs.

Once again the two shapes of curve present themselves, with points with low standard deviations in figure 6.3 showing a very different curve to the logarithmic curve seen by those in areas of high standard deviations.

6.3 Discussion

Results from figure 6.2 closely resemble those from figure 4.9 and support hypothesis 5. The fact that intervention algorithms show similar performance across different network structures is a strong indicator that success is due to the algorithms themselves and not an artefact of the network structure. Success on a network structure chosen to resemble real-world social networks also indicates that light-touch intervention methods are a useful focus of research for potential deployment on said real-world networks.

In third-space networks, where extremists do not initially control an agent with high (top quartile) degree, opinion-agnostic algorithms suffer a severe hit to their efficacy, as much as 20% fewer agents saved when compared with the same point in a normal scale-free network. Without additional information to differentiate between a large

number of similar agents with the same degree, opinion-agnostic algorithms are forced to select randomly. By contrast, opinion-aware algorithms are able to distinguish between similar agents using their opinions, and thus select appropriate and justified actions.

These results show that intervention methods are effective not only in random networks but also in those closer to reality. We chose to model two very different social networks to explore efficiency on mostly un-moderated forums where extreme beliefs are rewarded with page views, likes and comments, and on more moderate forums where extremists rarely rise to prominence. We find that in each case our newly proposed algorithms are able to identify and remove edges that are likely to lead to increased polarisation, particularly in cases where agents are very confident in their beliefs, and less uncertain.

In future work this aspect of the model could be improved through more accurate data in order to simulate a more realistic social network. A combination of psychological studies and automated linguistic analysis could yield realistic values for extremist proportions and uncertainty values that are close to reality, and thus worthy of further investigation (Fernandez and Alani, 2018).

6.4 Applying Interventions to Broadcast-Based Social Media

Rather than one-to-one conversations, information flow over social media typically presents as a one-way broadcast of information. Public replies and comments are often more like a replying broadcast than a single-target communication. To model this, we modified the update rule to only update the recipients of a message, while the sender remains unchanged in their opinion. This emulates the one-way nature of communication in a public social media post, and the lack of back-of-forth dialogue that is typical in a conversation. We further updated the group selection rule to allow the sender to

interact with all of their neighbours at once, to replicate the broadcast nature of public communication over social media. This modifies the standard Relative Agreement model to use the equations 6.1 through 6.3 rather than those used previously.

$$h_{ij} = \mathsf{min}(s_i + u_i, s_j + u_j) - \mathsf{max}(s_i - u_i, s_j - u_j) \tag{6.1}$$

$$s_i = s_i + \mu(s_j - s_i)\left(\frac{h_{ij}}{u_i} - 1\right) \tag{6.2}$$

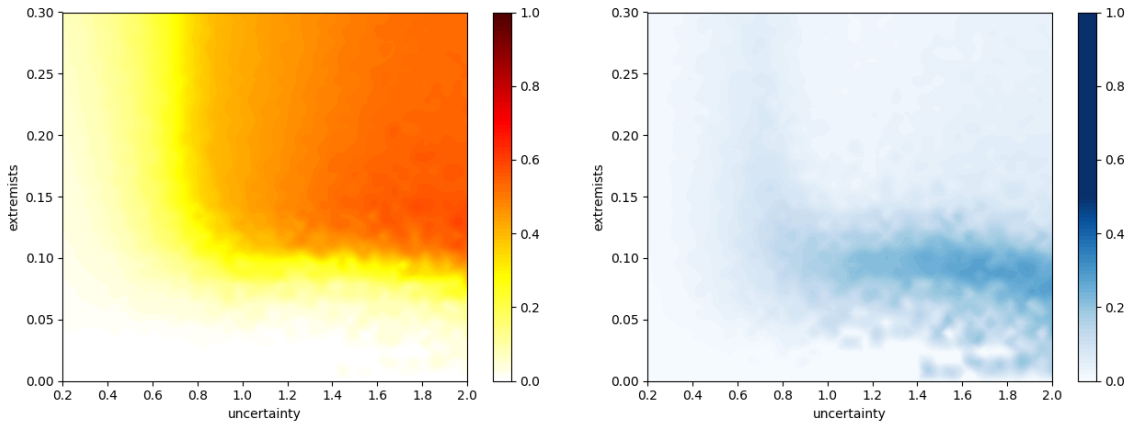
$$s_j = s_j \tag{6.3}$$

With these modifications, interactions become less like a conversation between agents and more like a public broadcast of opinion. All agents following that broadcaster are potentially then influenced, but the broadcaster themselves remains unaffected by posting a message.

6.5 Experimental Setup

Once again, experimentation followed the same procedure as discussed in section 4.2, with a random network. Furthermore, the update and group selection rules are altered. The update rule is changed to one in which the influencing agent is not influenced by those that they interact with, and the group selection rule is changed from selecting a pair of neighbours to instead selecting an agent and all of that agent's neighbours. We expect a reduced efficacy on this model, and so we alter our intervention methods to remove up to 25% of the edges within the network. We no longer count this as a "light-touch" intervention, but it allows us to see if there exists an effect at higher intervention levels. To allow for comparison with our work in previous chapters, we elect to continue using the grid pattern of points, with $x = 60, 100, 130, 160$ and $y = 0.1, 0.15, 0.2, 0.25$.

Test Case 7 *Intervention efficacy on this new broadcast-based model will be substantially*



(a) Mean **(b)** Standard Deviation

Figure 6.5: Heatmap for our modified RA model

reduced, as extremist influence will propagate much faster over the network.

6.6 Results and Discussion

As shown in figure 6.5, polarisation within this model becomes extremely strong as uncertainty rises, provided the proportion of extremists is above approximately 10%. However, it still bears a strong resemblance to the heatmaps shown in earlier chapters. This indicates that the changes to the model intensify the spreading of polarisation throughout the network in the absence of intervention strategies.

Within this model, despite earlier success in conversation-based simulations, opinion-agnostic intervention methods had a substantially reduced effect on reducing the spread of extremist opinions through the network. This demonstrates that alternative methods are required when dealing with broadcast-based communication than when dealing with conversation-based communication systems. Influence targeting, once surpassing a threshold of interventions, was able to significantly reduce extremism. This threshold is dependant on the average degree, as past this threshold targeted extremist agents have all edges removed which is functionally identical to deleting the agent.

Broadcast-based communications allow extremist agents a substantial increase in efficacy over one-to-one communications, in that they can persuade entire clusters of agents at once. With one-to-one communications, there is the potential for an influenced agent to be influenced back towards the centre by its peers. As one member of a cluster becomes more extreme, the rest of that cluster provide a countering force that influences the newly extreme agent back towards a central position. However with broadcast communication, those peers are themselves influenced by the extremist. This tendency to converge to the central opinion within a cluster is still present, only since all opinions are changed at once, the central opinion drastically changes. To prevent this, an extremist must be all but entirely removed from the network, which contradicts our aim of light-touch interventions. However, light-touch interventions are able to sufficiently insulate vulnerable individuals from being directly exposed to extremist agents.

However, these results and behaviours rely on the assumption that the only thing that can decrease uncertainty is interaction with a more confident individual. In this model, a large group of agents may all share the same opinion, but are still no more confident than they would be were they alone. In addition, all agents but extremists are equally uncertain. In such a system, only an agent interacting with an extremist or a previously radicalised agent can become more confident. While interactions with non-radical agents can bring a vulnerable agent's opinion back towards the centre, they cannot increase their confidence. Because of this, extremist interactions not only radicalise their victims, but make them more susceptible to future radicalisation and less susceptible to de-radicalisation. Non-extremist interactions can reverse the radicalisation, but with an efficiency decreasing over time during the simulation.

6.7 Conclusion

This section focused on scale-free networks, a type of graph frequently used to simulate graphs of social interactivity. We divided these into two types of network: one in which extreme views are prominent at the beginning of the simulation, and one in which those holding extremist views are prevented from having a degree in the top quartile.

In the former case, we find that our results match closely with those from a random network. However in the second case, the performance of our opinion-aware methods far outstripped the opinion-agnostic methods. A targeted approach therefore appears to be more useful in communities that already have some form of moderation in addition to edge removal that discourages extremist behaviour.

The methods described in chapter 4 found measurable success in reducing the spread of extremism and the trend toward polarisation within those networks. However, many social media communications present themselves as a one-way broadcast of information to all neighbours, rather than the conversations modelled so far. We have investigated our algorithm's efficiency on this communication model, and it is shown in our second experiment that this success is predicated on an unrealistic assumption: that agents only ever engage in one-to-one communication. As we have shown in this section and predicted in the specification of test case 7, when this assumption is discarded interventions short of effectively removing agents entirely from the network were less effective. This is particularly so for opinion-agnostic algorithms.

The highly interconnected nature of the networks we have used when combined with the broadcast communication module ensures that a message introduced at one node quickly spreads to virtually every other agent in short order, much like the multiple cascade models used in epidemiology. Under these conditions halting the spread of influence via connectivity-based means is ineffective as the message is all but guaranteed to reach every agent regardless.

Chapter 7

Conclusion and Future Work

Within this thesis, we have presented several contributions to the field of opinion dynamics, particularly that region of the field that deals with the spread of extremism, polarisation, and radicalisation within communities. The first of these is a unified framework for opinion dynamics that aims to bring together a number of disparate models. Approaching models in this modular manner not only speeds up replication of experiments but also allowed us to see differences and similarities across numerous authors and works. It also allows us to make changes to individual components of models with a minimum of effort and ensure no errors slip into other aspects of the model. We first utilise this capability to exchange the complete graph of the original Relative Agreement model with a random network structure.

Under these experiments, we find results that match with the conclusion of Amblard *et al.* in that single extreme convergence requires a critical connectivity (Amblard, Weisbuch, and Deffuant, 2003). As our chosen structures do not meet this critical threshold, we likewise do not observe single extreme convergence.

However, we view a dual extreme convergence state as also undesirable and so look to find a method of intervening on the network to prevent this. Dual extreme convergence indicates an extremely polarised political atmosphere, with agents unable to compromise

or even communicate with those holding different beliefs. This can lead to political instability or paralysis, conflict, and even physical violence in the most extreme cases.

However, the removal of agents from social networks *en masse* has several negatives: the network operators may be reluctant to eliminate portions of their user base; such interventions may be portrayed as censorship and stifling of discussion; extreme users may be useful for monitoring for intelligence purposes; and finally, the ease of creating a new account means that such measures are easily circumvented. For these reasons, we limited ourselves to the removal of edges: inhibiting the ability of extremist influencers to communicate directly with potentially vulnerable recruits.

We aimed to develop intervention strategies that can effectively target key edges that lead towards the polarisation of the network and remove them, while maintaining a light-touch presence on the network as a whole. We determined two broad categories of strategy, named "opinion-agnostic" and "opinion-aware". The former are strategies that solely take network-level information into account. Using prior metrics for network centrality, these algorithms determine the most central extremist agent and attempt to prevent them from pushing extremist influence into the network via severing their link to the most central non-extreme agent. Opinion-aware algorithms take more information into account regarding the precise opinions of vulnerable agents. With these strategies we place a stronger focus on those likely to become "infected" in the near future, and prevent them from propagating extremist influence into the network.

Under many conditions, we see a considerable increase in effectiveness using these new opinion-aware intervention methods. This increase is shown particularly strongly in areas of low uncertainty, whereas centrality-based methods have a slight advantage under conditions of high uncertainty.

At this point, our models had assumed that every edge was equal: that the cost of severing an edge was the same regardless of any other factors. However, this is not the case. Preventing an agent from receiving messages from a well-known and popular agent

is quite different from preventing them from hearing from a relatively unknown agent. Similarly, blocking a connection within a tightly-knit community with many mutual friends is different to blocking connections within a sparse community, or between different cliques. Therefore, our next step was to calculate costs for each edge, and repeat our experiments this time examining the reward/cost ratio rather than pure reward. However, this had very little effect on the polarisation and prevalence of extremism on the network as a whole. In most cases, factoring cost into the decisions merely altered the order in which edges were severed but did not change which edges were targeted. We did note however that some agents within real world social networks have added costs to intervening upon them, most notably political figures and organisations. These costs are not necessarily related to any network structural traits, and should be taken into account in any future work on this topic.

We then turned our attentions to the model itself. Thus far, the Relative Agreement model models interactions as conversations. Two agents interact at a time and may influence one another. However, we argue that a broadcast-based model may more accurately capture events on many of the major social networks in use today. In such a model, rather than communicating in a conversational manner, agents broadcast a message to all their peers at once and do not receive an immediate reply. This allows for a single agent to potentially influence many others without being influenced in turn - a marked difference from the Relative Agreement model. Under such a scenario we see similar trends in polarisation and extremism on the network, but meet with limited success in centrality-based interventions. Our targeted interventions are not severely affected, particularly Influence Targeting. This demonstrates to us the importance of modelling the interactions of agents accurately. What works for conversational-based communications does not work for broadcast-based communications, even when the method of adjusting an agent's opinion is kept the same.

It is made clear by our experiments that the uncertainty of agents is at least as

important to extremism as the proportionate presence of the extremists themselves. To explore this in more detail we present two further modifications to the Relative Agreement model in sections 7.1 and 7.2. In the first, rather than having all non-extreme agents given the same uncertainty, we assign them uncertainty based on a normal distribution. It is shown that this allows for confident centrist agents to counteract the influence of extremists to some extent, requiring a significantly higher average uncertainty than in the original model before dual extreme convergence is the norm. Our second modification changes the Relative Agreement update rule to allow interactions with similar agents to reduce an agent's uncertainty, and interactions with sufficiently dissimilar agents to increase uncertainty. Under this update rule, extremism is all but prevented from spreading through the network except under extreme circumstances.

The contributions of this broadcast-based model demonstrate the critical importance of having opinion formation models accurately reproduce the structure and rules of the situations they represent. While intervention methods built upon the original Relative Agreement model may find strong success there, they are unable to reduce polarisation in other environments. This is especially vital if conclusions drawn from such models are to be used in real-life applications.

In answering our research questions, we present several additional questions for future research. We chose to explore two opposite paradigms for target selection in our proposed algorithms: Vulnerability Targeting focused on vulnerable individuals with a goal of inoculating communities, while Influence Targeting focused entirely on vulnerable communities and discarded individual factors such as vulnerability to radicalisation. Future work could explore hybrid methods that exist in between these paradigms. We also limited our investigation into the effects of costing and value-judgements in interventions, and research into their effects and mechanisms of action on alternate network structures could be invaluable. Especially, as we note, pricing based on degree does not account well for structures with a high variance in degree distribution such as the scale-free

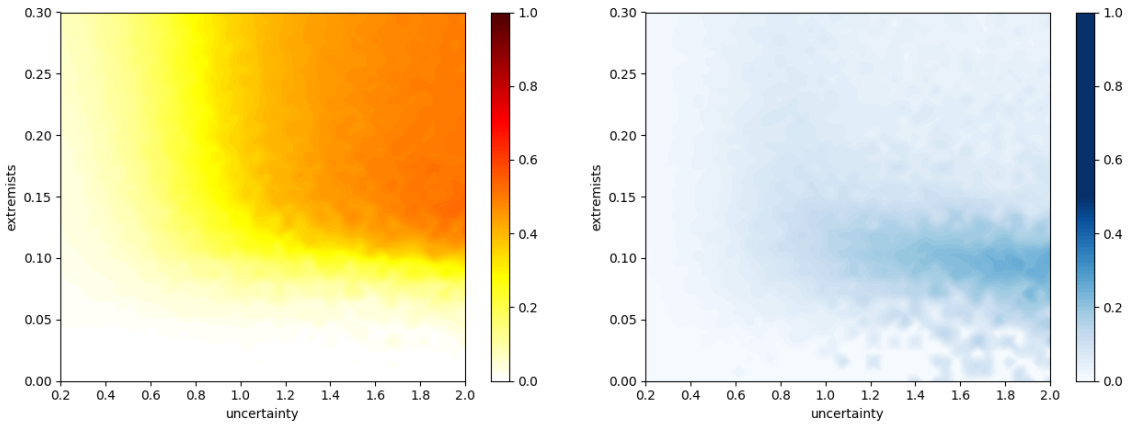
networks most frequently used in social network analysis. In further research we could also consider a deeper and more thorough exploration of this broadcast-based model, particularly into whether there exists an effective, computationally inexpensive, and light-touch intervention method that is able to reduce network polarisation using this new update rule.

In a similar manner as how this work drew upon ideas posed in the political, security, and epidemiological spheres, we hope that ideas posed within can be transferred back and inform developments there. In particular, we can consider the light-touch approach towards educating those hesitant about vaccination or other public safety safety measures, and in protecting the computer systems of users unaware of the importance of timely software updates. Using interventions similar to those posed in this work could allow medical and security professionals to focus their efforts where they are most beneficial, and to provide targeted interventions on patients or users who pose the greatest risk to themselves or their communities.

Below, we present preliminary findings for such future work investigating our broadcast model, wherein we alter the uncertainty parameter which had previously been left static. We propose normally distributed uncertainties where non-extreme agents have a chance to be more or less confident than others, and also a new update rule for modifying uncertainty. With this new rule, agents grow more confident through interacting with people similar to themselves.

7.1 Future Work: Normally Distributed Uncertainty

In this proposal for future work, we consider a variant of the Relative Agreement model that makes use of normally distributed values for uncertainty rather than having all non-extremist agents begin with the same uncertainty. This allows for non-extremist agents to begin more or less confident than others. We believe that this more accurately



(a) Mean **(b)** Standard Deviation

Figure 7.1: Heatmap for our modified RA model, with uncertainty under a normal distribution

represents real life, in that people hold both a wide range of opinions as well as a wide range in the strength of those convictions. Our preliminary results demonstrate a tentative step in this direction, and demonstrate the need for further investigation in this direction.

In figure 7.1, we recreated the heatmap using a normal distribution of uncertainties, rather than having all non-extreme agents have the same uncertainty. The normal distribution was created with the mean of the indicated value, and a standard deviation equal to $(x-0.1)/3$, where x is the mean uncertainty. As all extremists had an uncertainty of 0.1, this meant that an agent three standard deviations more confident than the mean was as confident as an extremist.

Test Case 8 *Extremism will be less prevalent with normally distributed uncertainties than with uniform uncertainty, as the more confident agents will influence others back to the centre.*

The heatmaps shown in figure 7.1 strongly resembles those shown in figure 6.5, except translated 0.2 to the right. Rather than beginning at $x = 0.8$, the large orange zone indicating total bipolar convergence begins at $x = 1.0$. This indicates that extremism requires a higher mean uncertainty to gain a foothold in a network with varying

uncertainties, and that the presence of abnormally highly confident moderate individuals more than balances out the effect of the presence of abnormally less confident agents. This confirms our hypothesis in 8. Future work could explore these boundaries within the heatmap, or explore the effects of other distributions for initial uncertainty.

7.2 Future Work: Decreasing Uncertainty With Homophilic Interactions

Four a second potential direction for future work, we consider a variant of the Relative Agreement model in which sufficiently similar agents with low confidence can increase their confidence through repeated interactions with one another. In this manner we emulate the "echo chamber" effect, where agents who are rarely if ever exposed to differing opinions become more and more resolute in their beliefs. Again, our results demonstrate a significant change in the model's behaviour that is worthy of future investigation.

In these experiments, we propose an adaptation to the rules governing uncertainty modification as a result of interactions under the Relative Agreement update rule. As written, interactions with a confident agent will increase your own confidence according to the overlap in opinions, as discussed in section 3.3. Our adaptation causes uncertainty to instead be reduced by 10% after an interaction with a similar agent, and increased by 10% after an interaction with a dissimilar agent. We supply two thresholds to determine similarity: an agent is deemed similar if their opinion lies within an assimilation threshold τ of one's own, and dissimilar if it lies further than a rejection threshold ϵ from one's own. This can be seen in equation 7.1.

$$u_i = \begin{cases} 0.9u_i & |s_i - s_j| < \tau \\ 1.1u_i & |s_i - s_j| > \epsilon \\ u_i & \text{otherwise} \end{cases} \tag{7.1}$$

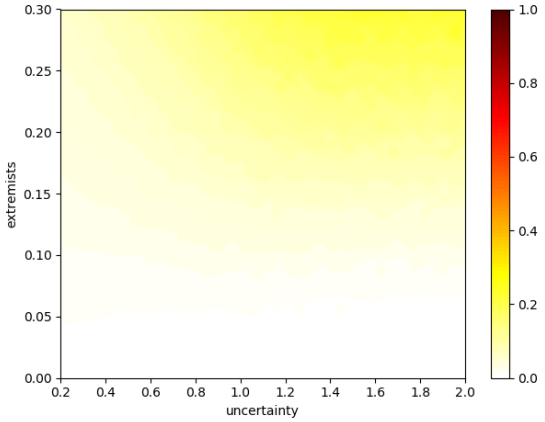
We predict that a central or moderate cluster will emerge, its size proportional to the assimilation threshold τ. This cluster will them reinforce its own confidence, and thus serve to attract new highly confident central agents. If this central cluster is sufficiently large, extremists will be unable to gather new recruits and thus the overall polarisation will be reduced.

Test Case 9 *When agents are sufficiently accepting, extremism will be drastically reduced.*

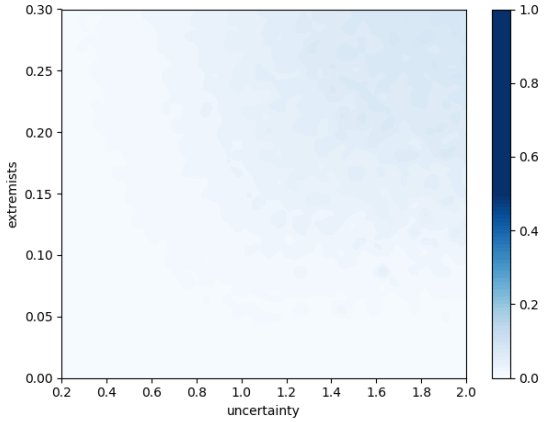
Figure 7.2 shows the dramatic effect this has on extremism throughout the network. For a relatively large assimilation threshold, even with a very high proportion of extremists initially within the network polarisation is kept to a relatively low level. For sufficiently low proportions of extremists this model even shows the elimination of extremism within the network, which was previously all but impossible. As we reduce the assimilation threshold however, the results begin to look strikingly similar to the original Relative Agreement model.

These results confirm hypothesis 9, but are by no means an exhaustive exploration of this new model. Rather, they serve to indicate that modifying the confidence updating mechanism of the Relative Agreement model gives drastically different results and is worthy of exploration in future work.

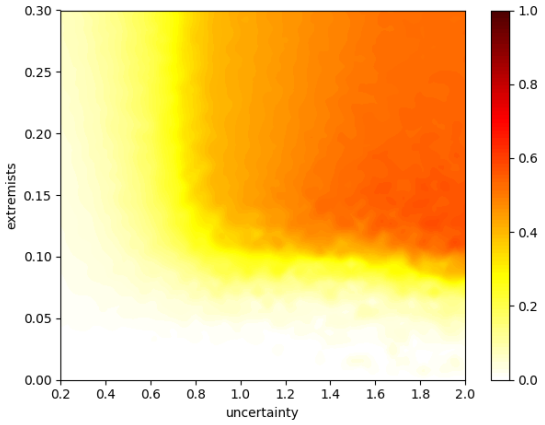
These initial findings for future work show that the behavioural changes made by these two relatively small alterations can have drastic consequences worthy of investigation, and also that our proposed framework and simulator are powerful tools for performing said investigation.



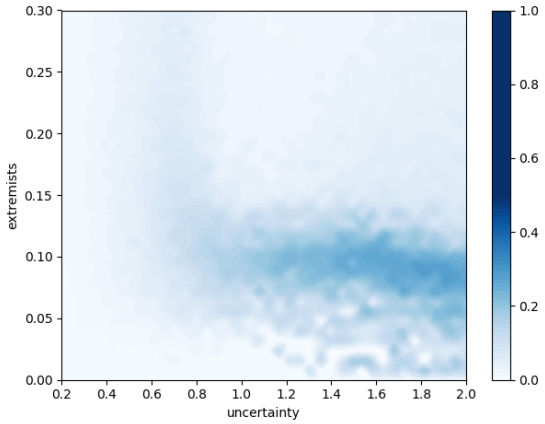
(a) Mean - assimilation 0.2, rejection 0.8



(b) Standard Deviation - assimilation 0.2, rejection 0.8



(c) Mean - assimilation 0.1, rejection 0.8



(d) Standard Deviation - assimilation 0.1, rejection 0.8

Figure 7.2: Heatmap for our modified RA model, with uncertainty reduced through interactions with similar agents, and increased through dissimilar agents

Bibliography

Alizadeh, Meysam (2012). "Essays on the Drivers of Political and Ideological Extremism". In:

Amblard, Frédéric and Guillaume Deffuant (2004). "The role of network topology on extremism propagation with the relative agreement opinion dynamics". In: *Physica A: Statistical Mechanics and its Applications*. DOI: 10.1016/j.physa.2004.06.102.

Amblard, Frédéric, Gérard Weisbuch, and Guillaume Deffuant (2003). "The drift to a single extreme appears only beyond a critical connectivity of the social networks Study of the relative agreement opinion dynamics on small world networks". In:

Axelrod, R (1997). "The dissemination of culture: A model with local convergence and global polarization". In: *Journal of conflict*. URL: https://journals.sagepub.com/doi/abs/10.1177/0022002797041002001.

Baldassarri, Delia and Andrew Gelman (Sept. 2008). "Partisans without Constraint: Political Polarization and Trends in American Public Opinion". In: *American Journal of Sociology* 114.2, pp. 408–446. ISSN: 0002-9602. DOI: 10.1086/590649. URL: http://www.journals.uchicago.edu/doi/10.1086/590649.

BBC News (Nov. 2019a). *Anonymous 'anti-Islamic State list' features Obama and BBC News*. URL: http://www.bbc.co.uk/newsbeat/article/34919781/anonymous-anti-islamic-state-list-features-obama-and-bbc-news.

— (Sept. 2019b). "Facebook will not fact-check politicians". In: *BBC News*. URL: https://www.bbc.com/news/technology-49827375.

BBC News (Jan. 2020). *Twitter sorry for letting adverts target neo-Nazis*. URL: https://www.bbc.co.uk/news/technology-51112238.

Bonacich, P (1987). "Power and centrality: A family of measures". In: *American Journal of Sociology*. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/228631.

Brandes, Ulrik and Daniel Fleischer (2005). "Centrality Measures Based on Current Flow". In: pp. 533–544. DOI: 10.1007/978-3-540-31856-9{_}44. URL: https://link.springer.com/chapter/10.1007/978-3-540-31856-9_44.

Castellano, Claudio, Santo Fortunato, and Vittorio Loreto (May 2009). "Statistical physics of social dynamics". In: *Reviews of Modern Physics* 81.2, pp. 591–646. ISSN: 0034-6861. DOI: 10.1103/RevModPhys.81.591. URL: https://link.aps.org/doi/10.1103/RevModPhys.81.591.

Castellano, Claudio, Vittorio Loreto, et al. (2005). "Comparison of voter and Glauber ordering dynamics on networks". In: *Phys. Rev. E* 71.6, p. 066107. arXiv: 0501599v1 [cond-mat]. URL: https://arxiv.org/pdf/cond-mat/0501599.pdf.

Chen, Duanbing et al. (2011). "Identifying influential nodes in complex networks". In: *Physica A*. DOI: 10.1016/j.physa.2011.09.017.

Clifford, Peter and Aidan Sudbury (1973). "A model for spatial conflict". In: *Biometrika* 60.3, pp. 581–588. DOI: 10.1093/biomet/60.3.581. URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/60.3.581.

Coaston, Jane (Aug. 2018). "Alex Jones banned from YouTube, Facebook, and Apple, explained". In: *Vox*. URL: https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-youtube-conspiracy-theories.

Coates, A, L Han, and A Kleerekoper (2018a). "A Unified Framework for Opinion Dynamics". In: *dl.acm.org*. URL: https://dl.acm.org/citation.cfm?id=3237857.

Coates, A, L Han, and A Kleerekoper (2018b). "A Unified Opinion Framework Simulator". In: *dl.acm.org*. URL: https://dl.acm.org/citation.cfm?id=3237983.

Concatto, F et al. (2017). "Genetic algorithm for epidemic mitigation by removing relationships". In: *dl.acm.org*. URL: https://dl.acm.org/citation.cfm?id=3071218.

Deffuant, Guillaume, Frédéric Amblard, et al. (2002). "How can extremism prevail? A study based on the relative agreement interaction model". In: *Journal of Artificial Societies and Social Simulation* 5.4.

Deffuant, Guillaume, David Neau, et al. (Jan. 2000). "Mixing beliefs among interacting agents". In: *Advances in Complex Systems* 03.01n04, pp. 87–98. ISSN: 0219-5259. DOI: 10.1142/S0219525900000078. URL: http://www.worldscientific.com/doi/abs/10.1142/S0219525900000078.

Deffuant, Guillaume, Gérard Weisbuch, et al. (2013). "The Results of Meadows and Cliff Are Wrong Because They Compute Indicator y Before Model Convergence". In: *Journal of Artificial Societies and Social Simulation* 16.1, p. 11. ISSN: 1460-7425. DOI: 10.18564/jasss.2211. URL: http://jasss.soc.surrey.ac.uk/16/1/11.html.

Erdős, Paul and Alfréd Rényi (1960). "On the evolution of random graphs". In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1, pp. 17–60.

Evans, Robert (2019). "Ignore the Poway synagogue shooter's manifesto: Pay attention to 8chan's/pol/board". In: *Bellingcat* 1.

Fan, Kangqi and Witold Pedrycz (2015). "Emergence and spread of extremist opinions". In: *Physica A: Statistical Mechanics and its Applications* 436, pp. 87–97. ISSN: 03784371. DOI: 10.1016/j.physa.2015.05.056.

Fernandez, M and H Alani (2018). "Contextual Semantics for Radicalisation Detection on Twitter". In: URL: http://oro.open.ac.uk/56501/.

Fernandez, M, M Asif, and H Alani (2018). "Understanding the Roots of Radicalisation on Twitter". In: URL: http://oro.open.ac.uk/id/eprint/54344.

First Amendment (Dec. 1791). URL: https://www.law.cornell.edu/constitution/first_amendment.

Fu, Feng and Long Wang (2008). "Coevolutionary dynamics of opinions and networks : From diversity to uniformity". In: *Phys. Rev. E* 78.1, p. 016104. DOI: 10.1103/PhysRevE.78.016104.

Gao, Bo, Zhenghong Deng, and Dawei Zhao (2016). "Competing spreading processes and immunization in multiplex networks". In: *Chaos, Solitons and Fractals* 93, pp. 175–181. DOI: 10.1016/j.chaos.2016.10.013. URL: https://www.sciencedirect.com/science/article/pii/S0960077916303083.

Gao, C et al. (2013). "Network immunization for interdependent networks". In: *Journal of Computational Information Systems* 9.16. ISSN: 15539105. DOI: 10.12733/jcisP0783. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.441.212&rep=rep1&type=pdf.

Gardner, Martin (1970). "MATHEMATICAL GAMES The fantastic combinations of John Conway's new solitaire game life". In: *Scientific American* 223, pp. 120–123. URL: http://ddi.cs.uni-potsdam.de/HyFISCH/Produzieren/lis_projekt/proj_ga....

Garimella, VRK (2017). "A long-term analysis of polarization on Twitter". In: *aaai.org*. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/viewPaper/15592.

Gil, Santiago and Damián H Zanette (2006). "Coevolution of agents and networks: Opinion spreading and community disconnection". In: arXiv: 0603295v1 [cond-mat].

Hadjichrysanthou, Christoforos and Kieran J. Sharkey (Jan. 2015). "Epidemic control analysis: Designing targeted intervention strategies against epidemics propagated on contact networks". In: *Journal of Theoretical Biology* 365, pp. 84–95. ISSN: 10958541.

DOI: 10.1016/j.jtbi.2014.10.006. URL: https://www.sciencedirect.com/science/article/pii/S0022519314005955.

Hagberg, Aric A., Daniel A. Schult, and Pieter J. Swart (2008). "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, pp. 11–15.

Hegselmann, Rainer and Ulrich Krause (2002). "Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation". In: *Journal of Artificial Societies and Social Simulation* 5.3. URL: http://www.math.fsu.edu/~dgalvis/journalclub/papers/02_05_2017.pdf.

Hine, Gabriel Emile et al. (2017). "Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web". In: *Eleventh International AAAI Conference on Web and Social Media*.

Holley, Richard and Thomas Liggett (1975). "Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model". In: *The Annals of Probability* 3.4, pp. 643–663.

Home Office (2018). *Prevent strategy*.

Hong, S and SH Kim Quarterly (2016). "Political polarization on twitter: Implications for the use of social media in digital governments". In: *Elsevier*. URL: https://www.sciencedirect.com/science/article/pii/S0740624X16300375.

Hosseini, Soodeh and Mohammad Abdollahi Azgomi (Oct. 2016). "A model for malware propagation in scale-free networks based on rumor spreading process". In: *Computer Networks* 108, pp. 97–107. ISSN: 13891286. DOI: 10.1016/j.comnet.2016.08.010. URL: https://www.sciencedirect.com/science/article/pii/S1389128616302523.

Khalil, E, B Dilkina, and L Song (2013). "CuttingEdge: Influence Minimization in Networks". In: *Workshop on Frontiers of Network Analysis: Methods, Models, and*

Applications at NIPS, pp. 1–13. URL: https://www.cc.gatech.edu/grads/e/ekhalil3/pdfs/CuttingEdge.pdf%20http://www.cc.gatech.edu/grads/e/ekhalil3/pdfs/CuttingEdge.pdf.

Kimura, Masahiro, Kazumi Saito, and Hiroshi Motoda (2009). "Blocking links to minimize contamination spread in a social network". In: *ACM Transactions on Knowledge Discovery from Data* 3.2, pp. 1–23. DOI: 10.1145/1514888.1514892. URL: https://dl.acm.org/citation.cfm?id=1514892.

Kozma, Balazs and Alain Barrat (2007). "Consensus formation on adaptive networks". In:

Krapivsky, P. L. and S. Redner (June 2003). "Dynamics of Majority Rule in Two-State Interacting Spin Systems". In: *Physical Review Letters* 90.23, p. 238701. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.90.238701. URL: http://link.aps.org/doi/10.1103/PhysRevLett.90.238701.

Liu, Yang et al. (July 2016). "A biologically inspired immunization strategy for network epidemiology". In: *Journal of Theoretical Biology* 400, pp. 92–102. DOI: 10.1016/j.jtbi.2016.04.018. URL: https://www.sciencedirect.com/science/article/pii/S0022519316300479.

Martins, ACR (2008). "Continuous opinions and discrete actions in opinion dynamics problems". In: *International Journal of Modern Physics*. URL: https://www.worldscientific.com/doi/abs/10.1142/s0129183108012339.

Martins, André C. R. and Cleber D. Kuba (Jan. 2009). "The Importance of Disagreeing: Contrarians and Extremism in the CODA model". In: *Advances in Complex Systems* 13.5, pp. 621–634. DOI: 10.1142/S0219525910002773. URL: http://arxiv.org/abs/0901.2737%20http://arxiv.org/abs/0901.2737.

Meadows, Michael and Dave Cliff (2012). "Reexamining the Relative Agreement Model of Opinion Dynamics". In: *Journal of Artificial Societies and Social Simulation* 15.4,

p. 4. ISSN: 1460-7425. DOI: 10.18564/jasss.2083. URL: http://jasss.soc.surrey.ac.uk/15/4/4.html.

Mendez, GR, AG Cosby, and SD Mohanty (2018). "Obamacare on Twitter: Online Political Participation and its Effects on Polarization". In: *researchgate.net*. URL: https://www.researchgate.net/profile/Gina_Rico_Mendez/publication/326326729_Obamacare_on_Twitter_Online_Political_Participation_and_its_Effects_on_Political_Polarisation/links/5b461324a6fdcc6619183829/Obamacare-on-Twitter-Online-Political-Participation-an.

Nandi, AK and H Medal (2016). "Methods for removing links in a network to minimize the spread of infections". In: *Elsevier*. URL: https://www.sciencedirect.com/science/article/pii/S030505481500249X.

Newman, M. E. J. (Jan. 2003). "The Structure and Function of Complex Networks". In: *SIAM Review* 45.2, pp. 167–256. ISSN: 0036-1445. DOI: 10.1137/S003614450342480. URL: http://epubs.siam.org/doi/10.1137/S003614450342480.

Newman, Mark E J, Stephanie Forrest, and Justin Balthrop (2002). "Email networks and the spread of computer viruses". In: *Physical Review E* 66.3, p. 35101.

Nonnenmacher, Vinicius et al. (2014). "Modeling and Visualization Individual and Collective Opinions towards Extremism in a Society". In: *Procedia Computer Science* 29, pp. 2412–2421. ISSN: 18770509. DOI: 10.1016/j.procs.2014.05.225. URL: http://linkinghub.elsevier.com/retrieve/pii/S1877050914004025.

OHCHR | International Covenant on Civil and Political Rights (Dec. 1966). URL: https://www.ohchr.org/EN/ProfessionalInterest/Pages/CCPR.aspx.

Pastor-Satorras, R and A Vespignani (2001). "Optimal immunization of complex networks". In: *Physical Review E* 65.3, p. 36104. URL: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.65.036104%20http://cdsweb.cern.ch/record/507290.

Quek, N (2019). "El-Paso Shootings: Growing Threat of White Supremacists". In: URL: https://dr.ntu.edu.sg/handle/10220/49921.

Richter, HLS (2019). "A Disunited Kingdom? Brexit, Social Media And The Rise Of Polarisation Through Echo Chambers". In: URL: https://dspace.library.uu.nl/bitstream/handle/1874/384434/Hannah%20Richter%20-%20Final%20Masters%20Thesis%20-%20A%20Disunited%20Kingdom.pdf?sequence=2.

Robinson, Katy, Ted Cohen, and Caroline Colijn (Mar. 2012). "The dynamics of sexual contact networks: effects on disease spread and control." In: *Theoretical population biology* 81.2, pp. 89–96. ISSN: 1096-0325. DOI: 10.1016/j.tpb.2011.12.009. URL: http://www.ncbi.nlm.nih.gov/pubmed/22248701%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3328800.

Rouchier, Juliette and Paola Tubaro (2011). "Can opinion be stable in an open network with hierarchy? An agent- based model of the Commercial Court of Paris". In: *The Effect of Information Technology in the Entrepreneurship (A Case Study in Golestan Province IRAN)* 10, pp. 123–131. ISSN: 1877-0428. DOI: 10.1016/j.sbspro.2011.01.015.

Santiago, R de, W Zunino, and F Concatto (2016). "A New Model and Heuristic for Infection Minimization by Cutting Relationships". In: *Springer*. URL: https://link.springer.com/chapter/10.1007/978-3-319-46672-9_56.

Szamrej, Jacek et al. (1990). "Andrzej From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact". In: *Psychological Review Ctyp* 97.3, pp. 362–376. URL: http://cognitrn.psych.indiana.edu/rgoldsto/complex/nowak90.pdf.

Sznajd-Weron, K and R Weron (2008). "Outflow dynamics in modeling oligopoly markets: The case of the mobile telecommunications market in Poland". In: *iopscience.iop.org*. URL: https://iopscience.iop.org/article/10.1088/1742-5468/2008/11/P11018/meta.

Sznajd-Weron, Katarzyna (Mar. 2005). "Sznajd model and its applications". In: URL: http://arxiv.org/abs/physics/0503239.

Sznajd-Weron, Katarzyna and Józef Sznajd (2001). "Opinion evolution in closed community". In: *International Journal of Modern Physics C* 11.6, pp. 1157–1165. arXiv: 0101130v2 [cond-mat].

Torregrosa, Javier and Ángel Panizo (2018). "RiskTrack: Assessing the Risk of Jihadi Radicalization on Twitter Using Linguistic Factors". In: pp. 15–20. DOI: 10.1007/978-3-030-03496-2{_}3. URL: http://link.springer.com/10.1007/978-3-030-03496-2_3.

Twitter (Nov. 2019). *Defining public interest on Twitter*. URL: https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html.

Universal Declaration of Human Rights (Dec. 1948). URL: https://www.un.org/en/universal-declaration-human-rights.

Urbig, Diemo, Jan Lorenz, and Herzberg (2008). "Opinion dynamics: The effect of the number of peers met at once". In: *Journal of Artificial Societies and Social Simulation* 11.4.

Vicario, M Del, F Zollo, and G Caldarelli (2017). "Mapping social dynamics on Facebook: The Brexit debate". In: *Elsevier*. URL: https://www.sciencedirect.com/science/article/pii/S0378873316304166.

Vice (July 2018). *Where Did the Concept of 'Shadow Banning' Come From?* URL: https://www.vice.com/en_us/article/a3q744/where-did-shadow-banning-come-from-trump-republicans-shadowbanned.

Webster, Lexi (2019). "Dissenter and Gab: the controversial platforms with implications for free speech". In: URL: https://theconversation.com/dissenter-and-gab-the-controversial-platforms-with-implications-for-free-speech-113392.

Weisbuch, Gérard, Guillaume Deffuant, and Frédéric Amblard (2004). "Persuasion dynamics". In: arXiv: 0410200v1 [cond-mat].

Wright, S, T Graham, and D Jackson (2017). "Third space and everyday online political talk: Deliberation, polarisation, avoidance". In: *eprints.whiterose.ac.uk*. URL: http://eprints.whiterose.ac.uk/119308/.

Xia, H, H Wang, and Z Xuan (2011). "Opinion dynamics: A multidisciplinary review and perspective on future research". In: *Journal of Knowledge and Systems Science* 2.4, pp. 72–91. URL: http://www.igi-global.com/article/international-journal-knowledge-systems-science/61135.

Appendix A

Additional Results

The following pages contain results for points chosen in a grid pattern, to better illustrate the efficacy of interventions as initial extremist proportion and uncertainty vary.

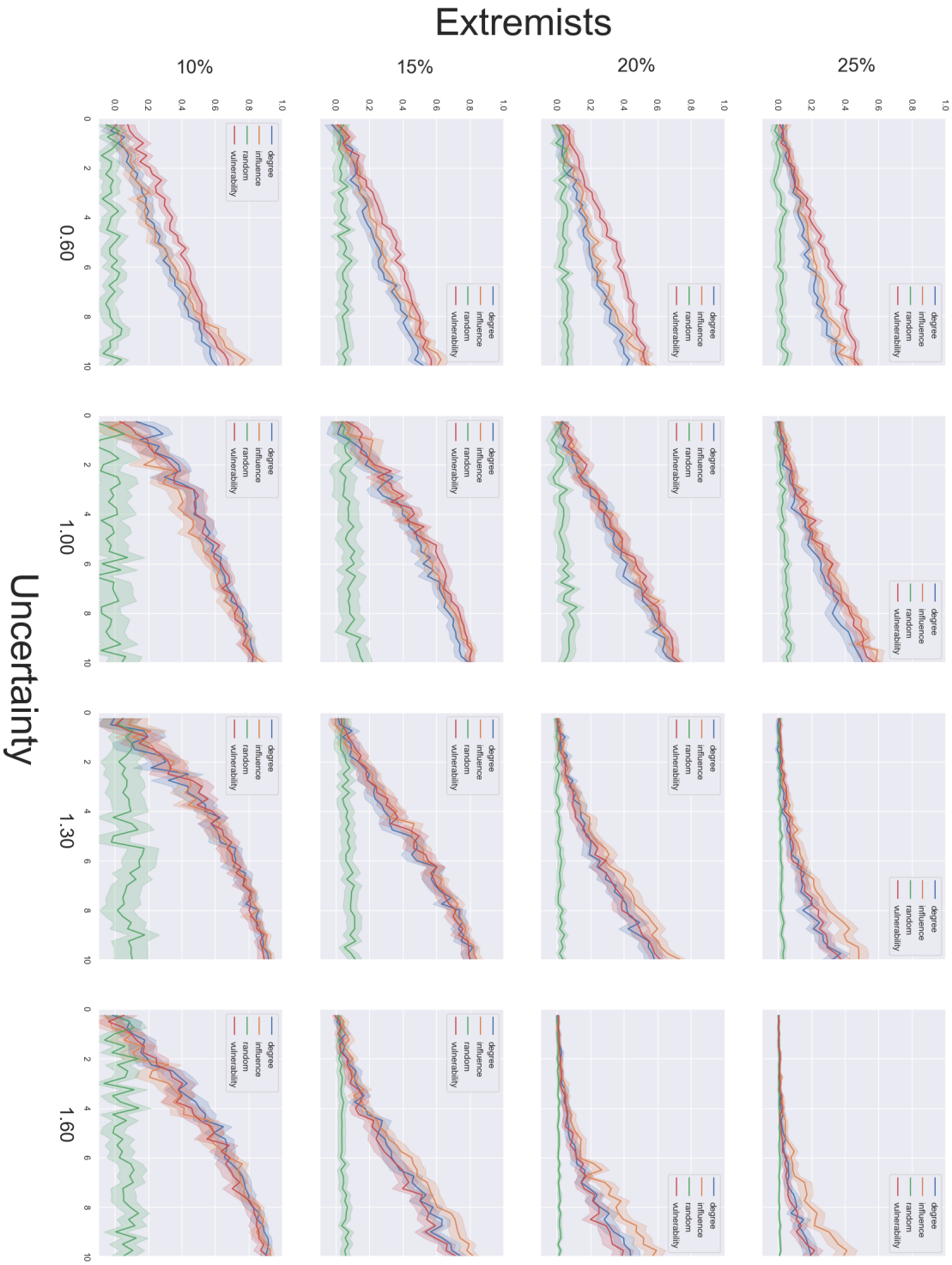


Figure A.1: Proportion of agents saved after up to 100 interventions.

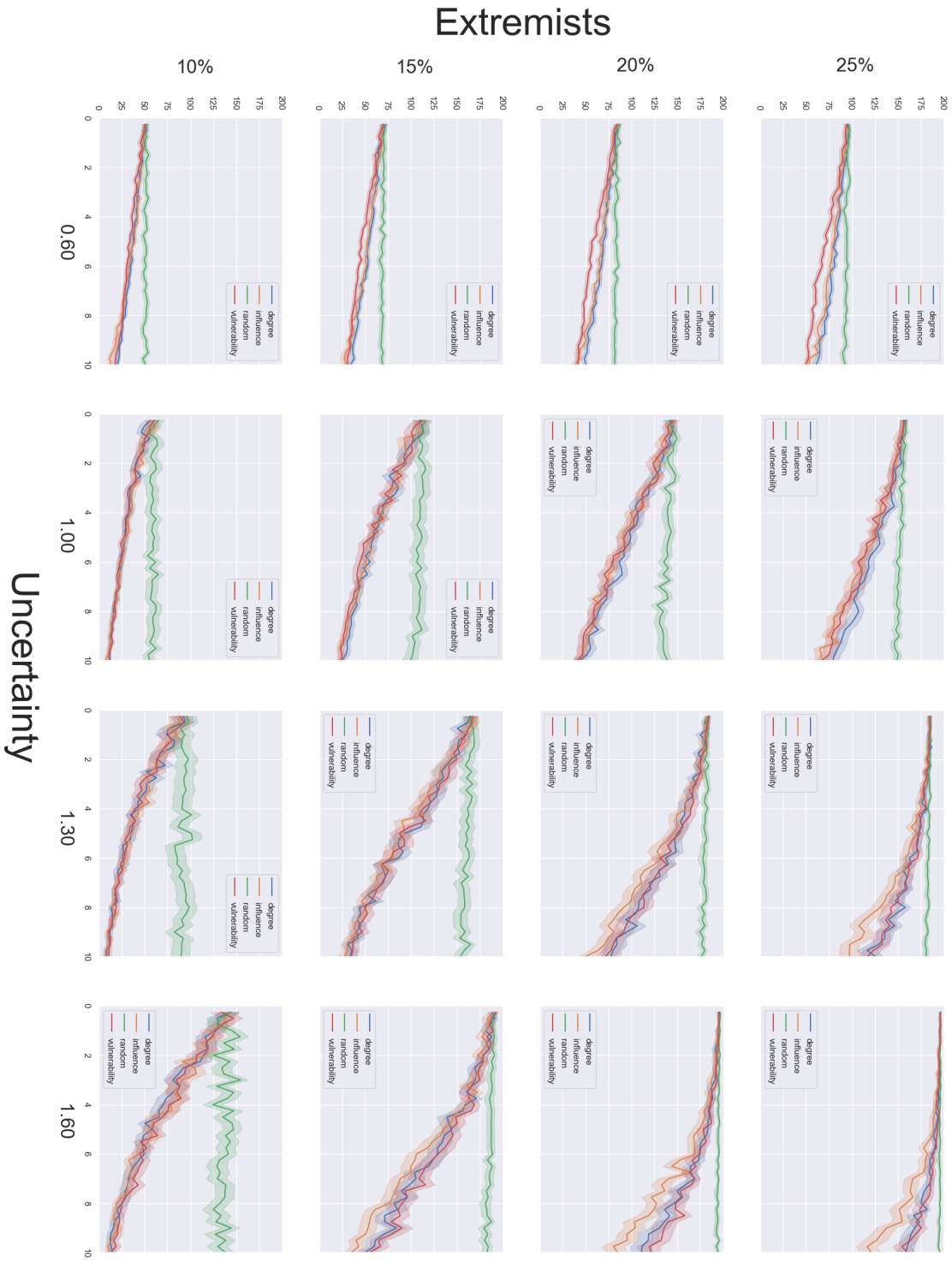


Figure A.2: Extremist agents remaining in the network after up to 40 interventions.

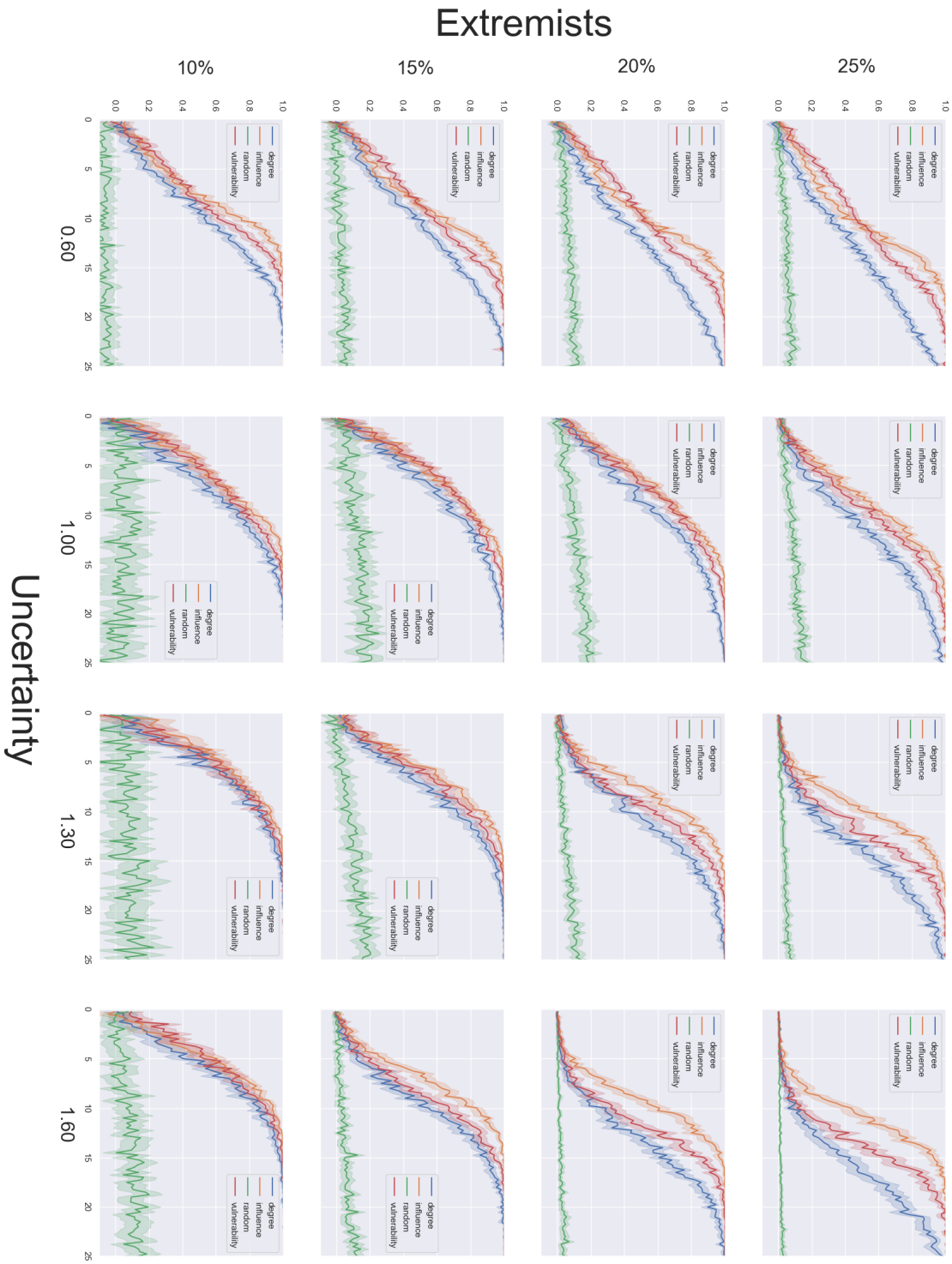


Figure A.3: Proportion of agents saved using the broadcast model after up to 100 interventions.

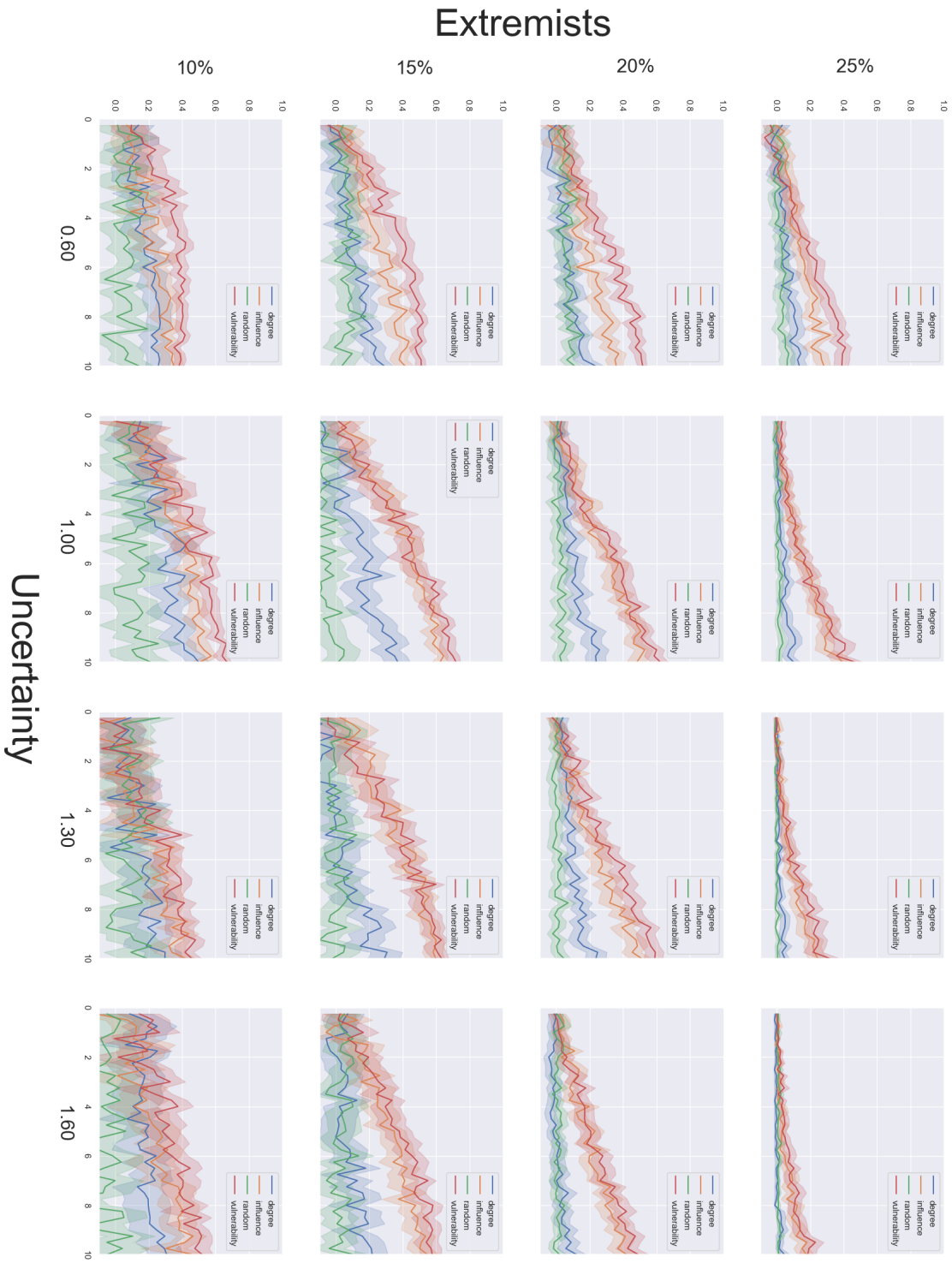


Figure A.4: Proportion of agents saved in the scale-free network without extremist hubs after up to 40 interventions.

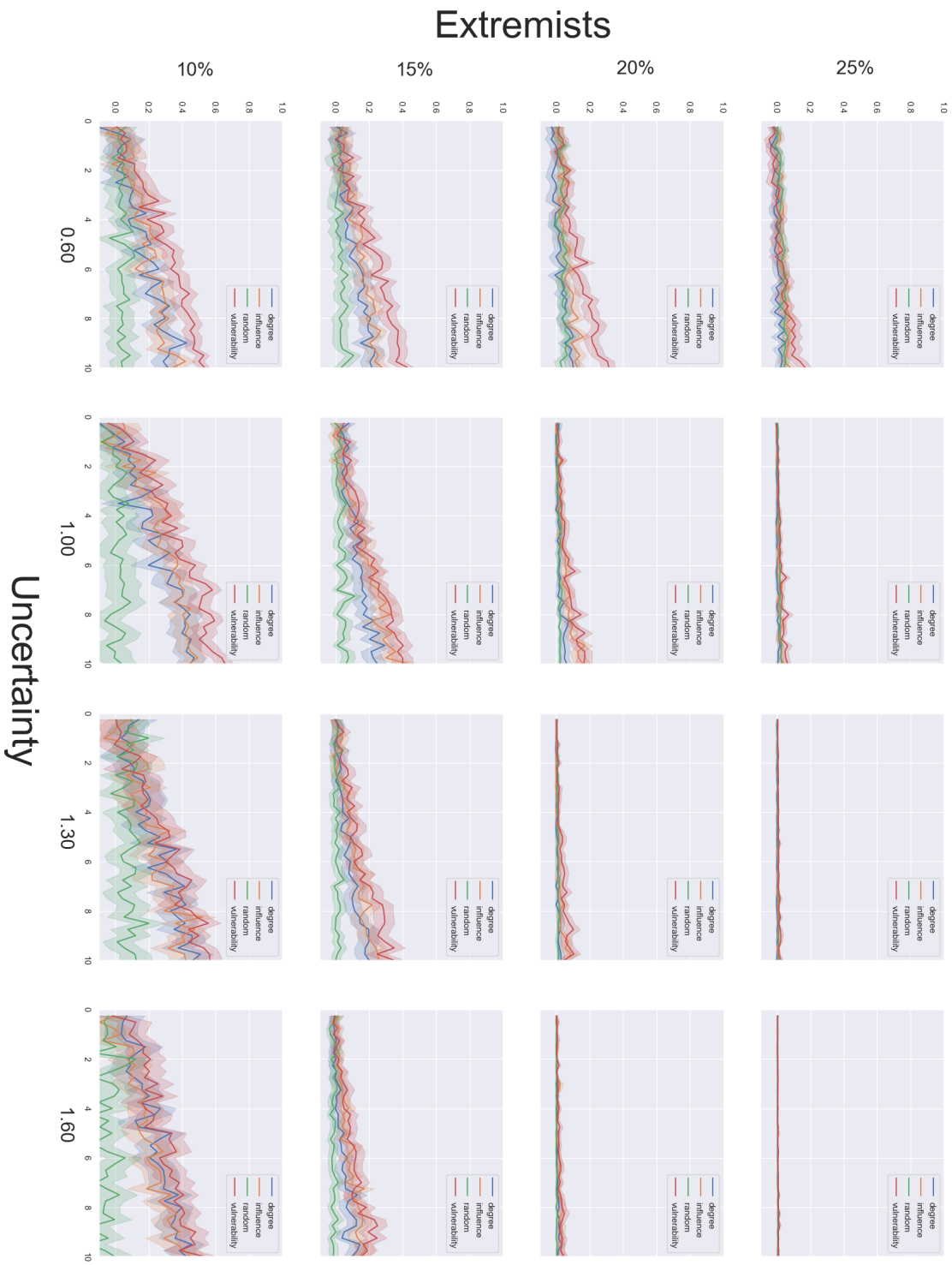


Figure A.5: Proportion of agents saved in the scale-free network with extremist hubs after up to 40 interventions.