Informatics

Auswirkungen von Interventionen in verschiedenen Meinungsbildungsmodellen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Data Science

eingereicht von

Anna Stonek, BSc
Matrikelnummer 11809899

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Dr. Stefan Neumann

Wien, 10. Oktober 2025

Anna Stonek

Stefan Neumann

TU WIEN Informatics

Assessing the Impact of Interventions in Different Opinion Formation Models

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Data Science

by

Anna Stonek, BSc
Registration Number 11809899

to the Faculty of Informatics

at the TU Wien

Advisor: Dr. Stefan Neumann

Vienna, October 10, 2025

Anna Stonek

Stefan Neumann

Erklärung zur Verfassung der Arbeit

Anna Stonek, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 10. Oktober 2025

Anna Stonek

Danksagung

Ich möchte mich zuallererst bei meiner Familie ganz herzlich bedanken, die mich während meines gesamten Studiums begleitet und immer unterstützt hat.

Des Weiteren möchte ich mich noch bei meinen Freunden Kateryna, Florian und Verena bedanken, die mich vor allem während des Schreibprozesses der Masterarbeit immer wieder ermuntert haben, weiter zu machen und immer ein offenes Ohr für meine Anliegen hatten.

Zusätzlich möchte ich mich noch bei meinen KollegInnen im Büro der Machine Learning Group der TU Wien bedanken, mit denen die Arbeit sehr viel Spaß gemacht hat.

Zu guter Letzt möchte ich noch meinem Betreuer, Dr.Stefan Neumann, meinen größten Dank aussprechen. Er hat sich jede Woche Zeit für ein Meeting genommen und hatte jederzeit einen Lösungsvorschlag für meine Fragen oder Probleme.

Acknowledgements

First of all, I want to thank my family for their constant support throughout all of my studies and specifically the process of writing my Master's thesis.

Moreover, I want to thank my friends Kateryna, Florian and Verena for their encouragement during the process of writing my thesis and for always having a friendly ear for my concerns.

Furthermore, I want to thank my colleagues at the office of the Machine Learning Group at TU Wien, whom I had a lot of fun working with.

Finally, I want to especially thank my supervisor, Dr. Stefan Neumann, who took the time to meet with me every single week and always had a suggested solution for my questions or concerns.

Kurzfassung

Aufgrund der immer größer werdenden Beliebtheit von Sozialen Medien wächst auch das Interesse von Widersachern, die Soziale Medien zu ihrem Vorteil nutzen wollen. Dies geschieht zum Bespiel durch die Verbreitung von Fake News oder Verschwörungstheorien. Der Meinungsbildungsprozess in sozialen Netzwerken kann durch Meinungsbildungsmodelle simuliert werden, wie zum Beispiel das FJ-Modell und das BC-Modell. Das Ziel dieser Diplomarbeit ist es, die Auswirkungen von Interventionen anhand von Polarisierung und Uneinigkeit in diesen zwei verschiedenen Meinungsbildungsmodellen zu untersuchen. Die Experimente haben ergeben, dass das FJ-Modell dazu tendiert, niedrigere Werte für Polarisierung zu erzeugen, als das BC-Modell. Betrachtet man ein bestimmtes Modell, so ist die Intervention, die mit diesem Modell berechnet wurde, am besten dazu geeignet, die Polarisierung in diesem Modell zu reduzieren. Werden allerdings mit anderen Modellen berechnete Interventionen in den Graphen eingefügt, dann steigen dadurch Polarisierung und Uneinigkeit nicht und so richtet dies keinen "Schaden" an. Zu guter Letzt können sogar gemischte Interventionen, die entsprechend der Beschreibung in dieser Arbeit konstruiert werden, die Polarisierung in sozialen Netzwerken effektiv reduzieren.

Abstract

Considering the widespread popularity of online social networks, it is only natural that adversaries have a rising interest in using those networks to their advantage, for example by spreading fake news or conspiracy theories. The opinion formation process in online social networks can be simulated with opinion formation models, such as the FJ-Model and the BC-Model. The goal of this thesis is to investigate the outcome of interventions in the two opinion formation models, and compare the results using polarization and disagreement. The experiments revealed that the FJ-Model tends to produce smaller values of polarization and disagreement than the BC-Model initially. Furthermore, considering one specific model, the intervention calculated with that model achieves the largest reduction of polarization. What is more, inserting an intervention calculated with a different model into the graph does not increase polarization and disagreement and therefore does not do “damage”. Finally, even mixed interventions constructed as described in this thesis can reduce polarization effectively.

Contents

CHAPTER 1

Introduction

In recent years, the popularity of online social networks only continues to grow. While some people use it as a means to connect with family and friends, others use it to promote media content. Evidently, social networks influence the opinions of their users, may that be through a discussion with their peers in the network, or an article or post they have read [TN22].

Although online social networks form an integral part of our everyday life nowadays, more and more experts start to point out the risks that come with social networking. Adversarial parties have started to use the opinion formation property of social networks to their advantage, for example by spreading fake news or conspiracy theories, as well as government agents trying to influence election results [DH21].

Therefore, the research interest in opinion formation models has grown immensely over the past years. Particularly, researchers are interested in understanding how social networks work, and how opinions are formed in this context. Additionally, the research interest on the impact of so-called interventions, for example dis- or misinformation campaigns, has grown as well. This research can help state agents and policy makers, to be able to apply appropriate measures when misinformation campaigns occur, to ensure societal stability [DH21].

In previous works, the impact of interventions in social networks has only been investigated on one specific opinion formation model [DH21, ZBZ21]. However, this only gives insight into the mechanics of that specific model. In this master's thesis, the impact of the choice of the opinion formation model will be investigated, and how the insights of the specific models are generalizable to other models.

The aim of this work is to investigate the impact of the choice of opinion formation model on the outcome of interventions in said model. As described above, the generalizability of the models will be investigated.

Two different opinion formation models, namely the popular Friedkin-Johnsen model [FJ90] and the bounded confidence model [HK02], will be implemented and consequently, an intervention on these two models will be modeled.

The first research question in this context will be "What is the impact of the choice of opinion formation model initially?" This question targets the results of the opinion formation models without any intervention.

The second research question will be "Given the intervention of inserting the same k edges into different models, what is the impact of the choice of opinion formation model?" This question targets the results of opinion formations models with a specific intervention, namely inserting the same k edges to every model.

And lastly, the third research question will be "When an algorithm tries to reduce polarization and disagreement across multiple models, to what extent does it still reduce polarization and disagreement?" This question targets whether an algorithm can be found that gives a sufficiently good solution across all models.

In Chapter 2, a comprehensive literature review of related work will be given. The origins and history of the FJ-Model and the BC-Model will be elaborated and discussed, as well as possible real-world applications of opinion formation processes. Furthermore, existing research on the outcome of interventions in both models will be presented and discussed.

In Chapter 3, the methodological approach to this thesis will be explained. First of all, the most important definitions around the FJ-Model and the BC-Model as well as the metrics used to evaluate the models, polarization and disagreement, will be given. Moreover, the datasets used for the simulations in this thesis will be described and the experimental setup will be explained. Lastly, the algorithms developed to find interventions of adding k edges to a social network will be presented.

In Chapter 4, the results of the experiments described in Chapter 3 will be presented. At first, the results of the experiments with models without any intervention will be presented. Subsequently, the results of polarization and disagreement with the interventions calculated for each model as well as the results with mixed interventions will be presented. Furthermore, the comparison of the results of the algorithms on the full search space and on the reduced search space will be reported.

And lastly, in Chapter 5, the results will be discussed and possible implications will be analyzed. The research questions will be answered and finally, conclusions from the analyses and discussions will be drawn.

CHAPTER 2

Literature Review

In this chapter, related work about the FJ-Model and the bounded confidence model, but also the outcome and impact of interventions will be analyzed and discussed. Moreover, a literature synthesis will be performed.

The scientific field of opinion dynamics originates from the social sciences and dates back to the 1950s, when early mathematical models were formulated in an effort to describe opinion formation dynamics in a group. French was particularly interested in how consensus can be reached [Fre56].

In this thesis, two particular models will be used for experiments: The Friedkin-Johnsen model and the bounded-confidence model. The first model used in this thesis, the Friedkin-Johnsen model was proposed by Friedkin and Johnsen in 1990 [FJ90]. This model will be formally introduced and analyzed below.

The second model used in this thesis is the bounded-confidence model, which was introduced by Krause in 1997 [Kra97], but thoroughly analyzed and described as the bounded-confidence model by Hegselmann and Krause in 2002 [HK02]. This model will be formally introduced and analyzed in the following section.

Lastly, related work on the outcome of interventions in social networks from both a mathematical and a social perspective will be discussed. More specifically, the UK's EU referendum will be used as an example to show the outcome of interventions from a social perspective.

2.1 The Bounded-Confidence Model

The bounded-confidence model was introduced by Krause in 1997 [Kra97] and further analyzed and described by Hegselmann and Krause in 2002 [HK02]. This model is based on the DeGroot model, which was developed in 1974 [DeG74].

The bounded confidence model is given by

$$z_i(t+1) = |I(i, z(t))|^{-1} \sum_{j \in I(i,z(t))} z_j(t)$$

for $t \in T = \{0, 1, 2, \dots\}$, where $I(i, z) = \{1 \leq j \leq n \mid |z_i - z_j| \leq \epsilon\} \cap E$ [HK02].

$z_j(t)$ is in our case the expressed opinion of a neighbor of node i, which is not more than ϵ different from $z_i(t)$, which is the expressed opinion of node i, at time point t. z_j is the expressed opinion of node j.

In a nutshell, agents in the bounded confidence model average over all neighboring agents with opinions that are not further away from their actual opinion than a given distance ϵ, the so-called "bound of confidence". The agents operate in a discrete-time context. This model was first introduced by Krause in 1997 [Kra97]. However, the term "bounded confidence" in reference to the model was described only one year later in 1998. The first comprehensive analytical and computational analysis of the model was then conducted by Hegselmann and Krause in 2002 [HK02].

The parameter ϵ is the so-called confidence level of the agents in the model. For simplicity purposes, a uniform confidence level is assumed for all agents. One particular feature of the model is that if a consensus exists, it will be reached in finite time. For an opinion profile $z = (z_1, \dots, z_n)$ there is a split between the agents i and j, if $|z_i - z_j| > \epsilon$.

An opinion profile $z = (z_1, ..., z_n)$ can be called an ϵ-profile if there exists an ordering $z_{i_1} \leq z_{i_2} \leq \cdots \leq z_{i_n}$ of the opinions such that two adjacent opinions are within confidence.

In a variant of the bounded confidence model, as published in [DNAW00], the updating of the opinions does not happen simultaneously, but it uses a pairwise sequential procedure. In order to distinguish the two models, albeit both are being called bounded confidence models, the models described above is called the Hegselmann-Krause model, and the variant just introduced is called the Deffuant-Weisbuch model. However, in this thesis, only the Hegselmann-Krause model is used.

Limitations In 2023, Hegselmann published another article on the bounded-confidence model, revisiting his analysis from 2002 and exploiting its limitations and misassumptions. He states to have overlooked a crucial feature of the model: For increasing values of the confidence bound ϵ, the analysis at the time suggested smooth transitions in the model's behavior. However, the transitions are actually wild, chaotic and non-monotonic [Heg23].

One example of this chaotic behavior is that the assumption of the initial analysis was that for increasing values of ϵ, the number of opinion clusters surviving after stabilization would decrease monotonically. But this assumption turned out to be wrong - there could very well be fewer final opinion clusters for increasing ϵ values, if the initial opinion distribution is held constant [Heg23].

Considering mis- and disinformation campaigns, the BC-Model in its extant form has an important limitation: All agents in the model are assumed to be benevolent, in a sense of never hiding their opinion from others in the group, let alone lying about their real opinions to mislead other group members and always being open to consider other agents's opinion. However, these assumptions are far from accurately depicting real social networks.

Furthermore, the confidence level ϵ is usually considered constant over time and the same for all agents in the model. In reality, people in a social network are not trusting towards other people to the same degree, nor will this degree of trust be established once and never change. On the contrary, people will most likely change their level of trust towards other people based on the experiences they make and the encounters they have [DH21].

Douven and Hegselmann [DH21] took some first steps to mending this situation by proposing two extensions to the BC-Model. In order to allow epistemically irresponsible agents in the model, they introduce new types of agents, where the irresponsibility can vary: The agents can either dogmatically stick to their opinion, and not let themselves be influenced by others, or they can not be interested in reaching a consensus at all.

Moreover, to mitigate the problem of fixed confidence levels, Douven and Hegselmann introduce the concept of confidence dynamics. They propose, that the level of confidence of a person is determined by the level of confidence of their peers. If someone is around broad-minded people, they themselves are likely to be broad-minded. If someone is around narrow-minded or suspicious people, they are likely to become, perhaps even unconsciously, narrow-minded themselves [DH21].

Extensions The BC-Model first presented in Krause [Kra97] has been widely used for investigating descriptive questions, more specifically, questions about the conditions that lead an initially disagreeing community to reach a consensus or, in the opposite case, to polarization [Lor08]. The focus of these studies lies on the time it takes to reach a stable final pattern [Kur15, KR11]. Other studies have used the model to investigate normative issues of interest mostly to philosophers, for example, the resolution of disagreement among peers [Dou10] and efficient truth approximation [DK11].

Many researchers also consider the BC-Model as a good starting point for their work, because it is relatively easy to extend or tweak the model to one's own needs. Hence, there are already many different flavors of the BC-Model out there. For example, Douven [Dou10] and De Langhe [DL13] present extensions of the model where agents receive "noisy" evidence. Crosscombe and Lawry [CL16] introduced an extension in which the opinion of an agent is represented by an interval rather than a number.

More complicated extensions can be found in Lorenz [Lor08], Jacobmeier [JAC05] and Pluchino et al. [PLR06], in which agents hold beliefs about multiple issues at the same time, instead of one issue at a time.

2.2 The Friedkin-Johnsen Model

The Friedkin-Johnsen model (FJ-Model for short) is a popular mathematical model for explaining and analyzing opinion formation processes in groups. It has even had practical applications, for example, the concatenated FJ-Model was used to explain the Paris Agreement negotiation process [BWV$^+$21] and the FJ-Model with multidimensional opinions was used to explain the change of public opinion in the US-Iraq war [FPTP16]. The FJ-Model is also the only opinion formation model that has been confirmed by a sustained line of human-subject experiments [BWV$^+$21, FPTP16].

In the FJ-Model, each node of the graph $G(V, E)$ holds an *innate opinion* $s_i \in [-1, 1]$ and an *expressed opinion* $z_i \in [-1, 1]$. The opinion formation process is modeled through iterative updating of the vector of expressed opinions, z. The update rule is given as:

$$z_i(t+1) = \frac{s_i + \sum_{j|(i,j)\in E} w_{ij} z_j(t)}{1 + \sum_{j|(i,j)\in E} w_{ij}}$$

for $t \in T = \{0, 1, 2, \ldots\}$.

The Friedkin-Johnsen Model is a very popular opinion formation model, originating in social sciences. It was meant to model the opinion formation process in a social network not as a static equilibrium, but rather as a dynamic process. One of the properties of the FJ-Model is the so-called innate opinions and the expressed opinions. The *innate opinions* are never changed during the opinion formation process, only the *expressed opinions*. So one could interpret the *innate opinions* as values or rather fixed beliefs of the users in a network, which are not easily changed. On the other hand, one could interpret the *expressed opinion* as the publicly shown opinion, which gets influenced by other users in the network [FJ90].

The opinion formation process in the FJ-Model is modeled through an update rule, as defined above. The opinions of users are being influenced by the opinions of neighboring users, suchlike users have a "friendship" with, for example. The interpretation of the edges of a graph representing a social network, specifically online social networks, may vary depending on the network. To sum up, the opinion of a user is the average of the opinions of neighboring users, also considering the innate opinion of the user. This behavior also manifests in colloquial speech, for example in the saying "You are the average of the people around you".

The FJ-Model has been studied extensively. In Proskurnikov et al. [PTCF17], a time-varying extension of the classic FJ-Model was developed. During the process of opinion formation, arcs of interpersonal influence may be added to or subtracted from the network and the influence weights assigned by an agent to their peers may be altered.

The *innate* opinions in the FJ-Model can also be interpreted as a prejudice that agents hold. The agents have some level of anchorage to their *innate* opinion and factor it into every opinion update. Unlike the bounded-confidence model [HK02], the FJ-Model describes the opinion formation process as linear discrete-time equations and is hence much simpler for mathematical analysis [PTCF17].

Moreover, the FJ-Model has been studied by experiments with real social groups, for example as described in Friedkin and Johnsen [FJ11] or in Friedkin et al. [FJB16].

In more recent works [PPTF17], necessary and sufficient conditions for the stability of the FJ-Model have been established. These conditions also provide convergence "on average" of its decentralized gossip-based counterpart [FRTI13, FIRT15].

Furthermore, multidimensional extensions of the FJ-Model have been developed, some of which have been used to describe the evolution of belief systems, representing agents' opinions on several mutually dependent issues [PTCF17, FPTP16, PPTF17]. The idea of incorporating multidimensional opinion vectors into the FJ-Model is employed to be more in line with reality. Previous research commonly focused on the opinion formation process over one single topic, but in reality, agents in a network often discuss several topics at a time. Hence, Zhou and Wu [ZW22] propose to extend the model with multidimensional opinion vectors and an increasing stubbornness parameter.

Other works have investigated interactions over random graphs using the FJ-Model with a community of stubborn agents [WXJ24] and a dynamic extension of the FJ-Model with the most general social network settings, which predicts that, under all possible interaction topologies, the emerging social power structures are determined by the agent's eigenvector centrality scores [JFB17].

Last but not least, there are some contributions made to the field of online social networks research by Neumann and Tu [TN22], which studies the interplay of opinion formation processes and information cascades in online social networks, and Musco et al. [MMT17], which studies the problem of minimizing polarization and disagreement simultaneously in online social networks. This line of work is specifically important for recommender systems: Should the system link and, therefore, exchange recommendations between two similar users to minimize disagreement or between two users with opposing viewpoints, to expose them to other views of the world and hence reduce polarization?

Another contribution made by Zhou et al. investigates the impact of timeline algorithms in the FJ-Model. The model is augmented by incorporating aggregate information from timeline algorithms, and studied to the aim of minimizing polarization and disagreement in the network [ZNGG24].

2.3 Interventions

The broad availability of fast and powerful computers has made agent-based modeling of social networks a popular tool for studying complex social phenomena that are difficult to investigate analytically. A relatively recent branch of this research field focuses on aspects of knowledge and belief acquisition, where the interaction of agents is central [DH21].

The paper [DH21] gives a good explanation of what an intervention in social networks is, including figurative examples to further clarify the meaning of an intervention. For this reason, [DH21] will be thoroughly discussed in the following paragraphs.

Intuitively, one could assume that the goal of a liar is to make others believe what they falsely assert; however, this is not always the case. Depending on the purpose of the liar, convincing others of what they falsely assert or diverting others from viewpoints that they tend to endorse, different strategies of deceit may be applicable [DH21].

Consider the following example for illustration: A politician is very convinced that their voter base will support them no matter what their view on issue X is (Let X be climate change). Their own voters do not care much about X; other voters who consider voting for the opponent, however, care a lot. The politician's strategy of deceit might now be to raise just enough doubt about the opponent's view on X among the voters to make them stay home on election day. To that end, the politician can lie more subtly ("Sea levels rise more slowly than scientists report") or more blatantly ("Climate scientists have been bribed by the liberal elites to publish results supporting green policies") about issue X. The strategy of the politician is hence not to convince voters of their own view of issue X, but to create just enough doubt in the opponent's view of issue X.

Accordingly, Douven and Hegselmann [DH21] define two types of campaigns: The misinformation campaign and the disinformation campaign. A misinformation campaign is an effort to deceive a target public about a given proposition X. The goal of a misinformation campaign is conversion, which means convincing the target public of a falsehood contrary to X.

A disinformation campaign is also an effort to deceive a target public about a given proposition X, but with the explicit goal of diversion. This means that a disinformation campaign aims to lure the target public away from believing X, but not necessarily convincing it of some other falsehood inconsistent with X.

Hence, a successful misinformation campaign always means a successful disinformation campaign. However, for this reason, a successful disinformation campaign is easier to realize than a successful misinformation campaign. It depends on the situation, whether a disinformation campaign is enough to achieve one's goals. In the example described above, the politician is running a disinformation campaign [DH21].

Similarly, it may be enough to win an election to suppress the enthusiasm of potential voters for the opponent. The voters might never vote for you, but just weakening the voter's turnout for the opponent might be enough to win the election. To mitigate the enthusiasm of potential voters it might be enough to divert the public from the truth or the opponent's view of certain issues, without the necessity of making them believe the falsehoods you are spreading, which makes this a disinformation campaign [DH21].

To investigate the influence of maleficent agents in the model, Douven and Hegselmann introduce a new type of agents, which are called the campaigners. These agents do not update their opinion at all, but keep their opinion fixed. That is, they do not let themselves be influenced by their peers.

During the experiments, it turns out that few campaigners manage better to convince the other agents of their opinion than a lot of campaigners. This seems counterintuitive, nevertheless it can be explained: The more campaigners there are, the more they pull the opinions of agents in their vicinity into their direction, which might result in an early split of the opinion vector, and then some agents stay out of reach for the campaigners. Fewer campaigners will also pull agents that are close-by into their direction, but not as strongly, and will therefore not provoke a split, which in turn means that they are more successful in spreading their campaign [DH21].

The second extension proposed by Douven and Hegselmann [DH21] is to add confidence dynamics to the bounded-confidence model. Initially, the confidence bound ϵ is set at a fixed value, equally for all agents in the model and it stays the same over all update steps. Realistically, however, people may want to adapt their broad- or narrow-mindedness over time, based on the experiences they make. So in this extended BC-Model, the confidence bound ϵ is subject to change and gets updated frequently. Based upon the update rule for an agent's opinion, which is heavily influenced by the opinion of their peers, the confidence bound is also updated by averaging over an agent's peer's confidence bounds.

This extension was also investigated as a defense mechanism against maleficent campaigners in the group. Campaigners do not let themselves be influenced by any other agent, which means that their ϵ is effectively 0. With the confidence bound dynamic, peers of the campaigners will update their respective ϵ to a smaller value, and narrowing their confidence bound. By becoming more selective in who to admit to one's peer group, agents can protect themselves from the influence of campaigners because they will more likely not appear in their peer group.

However, simply setting all ϵ boundaries to 0 is also not a solution, because this setting can annihilate the positive effects of social learning. In effect, the goal of an agent must be to find the right balance between being selective and accepting other agent's opinions.

To examine the optimization problem of reducing polarization and disagreement in online social networks from a mathematical perspective, Musco et al. [MMT17] and [ZBZ21] will be discussed.

Musco et al. investigated two approaches towards the minimization goal: The first approach examines the underlying graph topology, which enables or disables the minimization of polarization and disagreement, while the second approach examines the influence of the initial opinion profile on the minimization of polarization and disagreement [MMT17].

The optimal graph topology found by Musco et al. for their Twitter graph could reduce the polarization-disagreement index used as an objective function in their experiments by a factor of 6. The results were improved even further by reducing the number of edges in the optimal graph and transforming it into a sparse graph, which did not influence the polarization-disagreement index.

Even more impressive are the results for their Reddit data: The optimal graph topology was able to reduce the polarization-disagreement index by a factor of 60 000. The results were again further improved by reducing the number of edges in the graph and hence transforming it from a dense to a sparse graph.

One could assume that the structure of the optimal graphs had a strong community structure. Since there is always a tradeoff between polarization and disagreement, graphs with low disagreement would have two communities that are strongly connected within the communities and graphs with low polarization would have two communities that are strongly connected with each other. But in fact, the optimal graphs were more similar to random graphs [MMT17].

In [ZBZ21], the problem of minimizing polarization and disagreement via link recommendation is investigated. The authors show that their objective function is not submodular, although monotone. To tackle the exponential complexity of the combinatorial optimization problem, they resort to a greedy algorithm by iteratively adding the most promising edges.

In order to account for scalability for larger graphs, they also propose a faster greedy algorithm, where the computation time is significantly reduced compared to the naïve greedy algorithm. Seven different methods of finding interventions (a set of k edges that reduces polarization and disagreement) were compared: Selecting the optimal edge set found by exhaustive search, selecting k edges at random, selecting the top-k edges in terms of betweenness [Bra04], selecting the top-k edges with the largest product of two end-points degrees, selecting the top-k edges with the largest sum of two end-points degrees, selecting the edge set found by the naïve greedy algorithm and lastly, selecting the edge set found by the fast greedy algorithm.

During experiments, it turned out that the naïve greedy algorithm consistently outperformed all other methods, while the fast greedy algorithm yielded a close approximation of the naïve greedy algorithm and outperformed the remaining four methods on all datasets.

2.3.1 Examples of Interventions during the UK's EU referendum

However, considerable efforts were made during the EU referendum in 2016 to convince the British public that the option to leave the European Union was the right choice to make. According to the report of the Digital, Culture, Media and Sports Committee of the House of Commons of the British Parliament, there were several companies involved in data analytics and the processing of personal data of UK citizens in order to create targeted political advertisements [DC19].

At the center of these political advert campaigns was Aggregate IQ, a Canadian digital advertising web and software development company founded in 2012 by its owners Jeff

Silvester and Zack Massingham. AIQ carried out online advertising work for Brexit-supporting organisations such as Vote Leave, Veterans for Britain, Be Leave and DUP Vote to Leave [DC19].

To illustrate, a competition was held among football fans to predict the results of the European football championships, offering to win £50 million, which was in fact a data-harvesting initiative run by AIQ for Vote Leave. To enter the competition, fans had to enter their name, address, email, and telephone number, and additionally an indication of how they would vote in the EU referendum. AIQ processed all data gathered through this competition and harvested Facebook IDs and emails from the contest [DC19].

Furthermore, AIQ used three machine learning pipelines to identify people's photographs, match them to their individual Facebook profiles, and target advertising at these profiles [DC19].

Moreover, there is a technology called Facebook pixels, which is a plugin for websites and allows to track which Facebook users visited the website. In turn, Facebook can use the information gathered by the pixel about the users who visited the site, to allow advertisers to target said Facebook users. If a user visited a website with a pixel placed there by AIQ/Vote Leave, they could unknowingly be served adverts by that campaign through Facebook [DC19].

Additionally, Facebook users could also be served adverts from other campaigns if those had access to the same pixel data. Given that all of the campaigns mentioned above - Vote Leave, Be Leave, Veterans for Britain and DUP Vote to Leave - were managed and carried out by AIQ, it is reasonable to assume that all of the harvested data were synchronized and used for targeted advertising [DC19].

CHAPTER 3

Methodology

In this chapter, a complete overview of the methods used to create the experiments run for this master's thesis will be given. First of all, definitions of the mathematical models and the metrics used to evaluate said models, polarization and disagreement, will be given. Subsequently, the datasets used to run the simulations in this thesis will be described and the generation process will be explained. Moreover, the setup for the experiments run in this thesis will be explained. Lastly, the greedy algorithms developed to find interventions of adding k edges to a social network will be described.

3.1 Definitions

First of all, the models used in this thesis have to be defined, as well as the performance metrics, on which the models will be evaluated.

Let $G(V, E)$ be a connected graph with $n = |V|$ vertices and $m = |E|$ edges. We set $L = D - A$ to the Laplacian of G, where D is the diagonal matrix with $D_{i,i} = \sum_{j:(i,j)\in E} 1$ and A is the adjacency matrix with

$$A_{i,j} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{else} \end{cases}$$

We write I to denote the identity matrix and the dimension will typically be clear from the context.

3.1.1 FJ-Model

Recall the definition of the FJ-Model: Each node maintains a so-called *innate opinion* s_i that remains constant, and a so-called *expressed opinion* z_i that gets updated in every time step according to the following rule:

$$z_i(t+1) = \frac{s_i + \sum_{j|(i,j)\in E} w_{ij} z_j(t)}{1 + \sum_{j|(i,j)\in E} w_{ij}}$$

The innate opinions are in the interval [-1,1], i.e. $\sum_{i\in V} s_i = 0$ and $s_i \in [-1,1]$ for all $i \in V$. This implies that also $z_i(t) \in [-1,1]$. These assumptions are made without loss of generality as they can always be achieved by rescaling the vector s.

Commonly, the FJ-Model is defined in a weighted version, using the weight matrix $w_{i,j}$. However, in the case of this thesis, these weights are all assumed to be 1. There are other works using the weighted version, but in this thesis, it will not be considered.

Because the FJ-Model tends to produce very low values of polarization and disagreement for some datasets, the so called *stubbornness-parameter* c was introduced, in order to achieve meaningful results and, subsequently, to be able to actually reduce polarization and disagreement.

The update rule with stubbornness-parameter c is as follows:

$$z_i(t+1) = \frac{c\, s_i + \sum_{j|(i,j)\in E} w_{ij} z_j(t)}{c + \sum_{j|(i,j)\in E} w_{ij}}$$

Let z^* be the equilibrium of this process, meaning that the opinion profile z meets the convergence criterion defined below. The value z_i^* is the *expressed opinion* of node i. This equilibrium $z*$ is the solution to a linear system of equations:

$$z^* = lim_{t\to\infty} z(t) = (I+L)^{-1}s$$

I stands for the identity matrix and L is the Laplacian matrix of the connection graph G, as defined above. Most importantly, $I + L$ is always invertible because it is positive definite.

3.1.2 BC-Model

Recall the definition of the BC-Model: Each node updates its opinion by averaging over the opinions of only those neighboring nodes, whose opinions are not more than ϵ different from their own opinion. All other nodes are not considered for the opinion update of node i.

The update rule for the opinion of the i-th node z_i is therefore given by:

$$z_i(t+1) = |I(i, z(t))|^{-1} \sum_{j \in I(i,z(t))} z_j(t)$$

where $I(i, z) = \{1 \leq j \leq n \mid |z_i - z_j| \leq \epsilon\} \cap E$

Notice that in this model, there is no differentiation between an *innate opinion* and a *expressed opinion*, as opposed to the FJ-Model. In the BC-Model, there exists only one type of opinion, which is the expressed opinion.

3.1.3 Metrics

In the following paragraphs, the two metrics used to analyze and compare the FJ-Model and the BC-Model will be defined and explained. For the FJ-Model, let $z^* = (I + L)^{-1}s$ be the equilibrium vector of opinions according to the FJ-Model, for a social network $G(V, E)$ and innate opinions $s : V \to [-1, 1]$.

Polarization

The polarization should intuitively measure how far opinions deviate from the mean of opinions (the average). The mean of a vector of opinions z is defined as $\overline{z} = \frac{1}{n} \sum_{i \in V} z_i$. The polarization is then defined by:

$$P_{G,s} = \sum_{i \in V} \frac{(z_i^* - \overline{z})^2}{n}$$

Polarization is normalized by the number of nodes n in the network.

Disagreement

The disagreement of a social network is the squared difference between the opinions of i, j at equilibrium, normalized by the number of edges m in the network:

$$D_{G,s} = \sum_{(i,j) \in E} \frac{w_{ij}(z_i - z_j)^2}{m}$$

3.1.4 Intervention

An intervention is a set of k edges that is inserted into the given set of edges E of the graph $G(V, E)$.

This can heavily influence the outcome of a model regarding polarization and disagreement, since purposefully added edges can increase or decrease polarization and disagreement.

3.2 Datasets

In this section, on overview of the datasets used for the experiments will be given. For each dataset, the characteristics of the graph will be described and for the constructed datasets, a detailed description of the construction process will be given.

Each dataset consists of one graph G given by its vertices and edges and one vector of inital opinions, which is kept constant. For the BC-Model, these opinions represent the initial opinion profile and for the FJ-Model, the opinions represent the innate opinions. Each opinion is a real number in the interval $[-1, 1]$.

3.2.1 Twitter Large

The dataset Twitter Large was obtained from the repository of the paper [ZNGG24]. The graph consists of $n = 27058$ nodes and $m = 268860$ edges. The data was collected from a list of $\mathbb{X}$ accounts who actively engage in political discussions in the US, for which the entire list of followers up until a number of 100,000 followers was obtained. The data collection process was started in March 2022 and lasted one week [ZNGG24].

The vector s of initial opinions was also obtained from the repository of the paper [ZNGG24], which contains opinions as real numbers in the range of -1 and 1. The graph obtained from the repository is undirected, but in order to compare the undirected and directed case on the same dataset, a directed version of Twitter Large was created from the existing dataset.

3.2.2 Reddit

This dataset was obtained from the repository of the paper [CR20]. The graph consists of $n = 556$ nodes and $m = 9431$ edges. The nodes represent users and there is an edge between two users if there exist two subreddits (other than politics) that both users posted in. The topic of interest is politics.

The vector s of initial opinions was also obtained from the repository of the paper [CR20]. The graph obtained from the repository is undirected and to compare the undirected and directed case, a directed version of this dataset was created from the existing graph.

3.2.3 Flixster

The Flixster friendship network dataset was obtained from the network catalogue and repository Netzschleuder. The graph contains $n = 2523386$ nodes and $m = 7918801$ edges and it is undirected. The data set represents a network of friendships among Flixster users, an online social movie website that allows users to share movie ratings and connect with each other. Nodes represent users and an edge exists if two users are friends [J.16].

As mentioned above, the original graph as obtained from Netzschleuder is undirected, but for the purpose of comparing the undirected and directed case, a directed version of the dataset was created from the original one. What is more, the vector s of initial opinions was created by drawing from a uniform distribution in the interval of $[-1, 1]$.

3.2.4 Stochastic Block Model

This dataset was constructed according to the stochastic block model. The distinctive feature of an SBM graph are the communities in the graph.

The graph is constructed by creating n nodes, that form the vertex set V. This vertex set is divided into 3 subsets $V_1, V2, V_3$, the so-called communities. V_1 contains the first 200 nodes of V, V_2 contains nodes $n_{201} - n_{400}$ of V and V_3 contains nodes $n_{401} - n_{500}$. Subsequently, the nodes of the graph are structured into 3 communities, two of which contain 200 nodes and one of which contains 100 nodes.

The edges are then added according to the following rule: There are two parameters p and q, which are probabilities. With probability p, a new edge will be added within a community, and with probability q, a new edge will be added to a different community. This means that the resulting graph has strongly connected communities, while the communities among themselves are only loosely connected.

With these SBM graphs, the aim was to model social groups within one social network, which can often be observed also in real life. Not everybody is connected to every other user in the same way, but most often interest groups about the same topics arise, which is reproduced in the model.

Similarly, the initial opinion profile can be constructed in different ways. At first, the initial opinion profile was constructed using a uniform distribution, creating real random numbers in the interval $[-1, 1]$. However, due to the construction of the graph with strongly connected communities, the initial opinion profile can also be created using the community structure. In this second variant, each community is given a different mean of opinions, to which a certain noise is added. For example, if the SBM graph has 3 different communities, the means of the communities could be chosen as -0.5, 0 and 0.5 and the noise could be chosen in the range of -0.25 and 0.25, to let the opinions stay within the boundaries of -1 and 1. This way, the initialization of the opinions can reflect different communities having different opinions.

The graphs were created using the NetworkX package for Python. In total, two SBM graphs were created, to investigate the impact of the parameters p and q on the opinion formation process. The graph SBM High Cohesion has $n = 500$ nodes and $m = 47970$ edges and it was constructed with the parameters $p = 0.5$ and $q = 0.02$, whereas the graph SBM Low Cohesion has also $n = 500$ nodes and $m = 10624$ edges and it was constructed with the parameters $p = 0.1$ and $q = 0.01$.

Both graphs were initially created as directed graphs and, for the purpose of comparing the undirected as directed cases, an undirected version of the graphs was created from the original one by omitting the edge directions.

3.2.5 Barabasi-Albert-Graphs

These datasets are are constructed according to the Barabasi-Albert-Graph model [AB02]. The distinctive feature of this graph is that additional edges will preferably added to nodes with a high degree.

The graph is built as follows: In the beginning, m nodes are created, all of which are connected with all other nodes. Subsequently, every newly added node i is connected to m already existing nodes in the network. The connection takes places stochastically - the probability of connecting with an already existing node is proportional to its total number of connections. In other words: the more connections a node already has, the more likely it is to get new ones [JAC05].

This mimics the behavior of online social networks, where nodes with a high degree could represent influencers, who have a lot of followers. Newly added nodes could represent new users, which are likely to follow influencers, i.e. in the graph create an edge to a node with a high degree. The parameter n depicts how many nodes the graph has and m determines how many edges each newly added node adds to the graph.

Due to this, three Barabasi-Albert graphs were created using the NetworkX package in Python using different parameter configurations. The first graph has $n = 500$ nodes and $x = 4900$ edges and it was constructed with the parameter $m = 10$. The second Barabasi-Albert graph also has $n = 500$ nodes and $x = 9600$ edges and it was constructed with the parameter $m = 20$, whereas the last Barabasi-Albert graph also has $n = 500$ nodes and $x = 14100$ edges and it was constructed with the parameter $m = 30$.

All three of these graphs were initially created as directed networks, and for the purpose of comparing the undirected and directed case, an undirected version of the networks was created by omitting the edge directions.

3.3 Experimental Setup

All experiments of this thesis were run on a server of the Machine Learning Research Group of TU Wien. This is an ASUS Server with 2 AMD EPYC 9124 16-Core processors and 512GB RAM. The implementation can be found in the following GitHub repository: `https://github.com/11809899/comparison-of-opinion-formation-models`.

Convergence

The criterion for convergence of a social network in this thesis is defined by:

$$||z(t+1) - z(t)||_2 < 0.01$$

Model implementation The experimental setup first included the implementation of both the FJ-Model and the BC-Model. For each dataset described above, in total

5 models were run until the convergence criterion specified above was reached: the FJ-Model as well as the BC-Model with 4 different configurations of the ϵ-parameter, more specifically 0.05, 0.2, 0.5 and 0.75.

All of these models were run once on the directed version of the dataset and once on the undirected version of the dataset. Finally, the polarization and disagreement of the resulting opinion profile z^* was calculated and the opinion profile was stored.

The goal of this process was to obtain initial results of polarization and disagreement of the models without any intervention and to create a baseline against which later results could be compared.

Due to the very low values obtained for polarization and disagreement in the BC-Model with $\epsilon = 0.75$, this model was excluded from further analysis.

Finding interventions Furthermore, experiments were conducted to find interventions to reduce polarization in social networks. The goal of interventions is to find a set of edges that, once inserted into a network, reduce the overall polarization in the network.

There were two greedy algorithms implemented: One algorithm tries to find the intervention on the full search space of the graph, specifically the full complementary graph, which, depending on the network, can be quite large and the search for the greedy optimal intervention can be computationally and time-wise very expensive.

Hence, an approximation algorithm was implemented: The nodes with the greatest and the smallest opinions were collected to form candidate nodes, between which the edges for the intervention are searched. These potential edges therefore form a reduced search space compared to the full search space of the whole complementary graph, which is computationally much more efficient.

The resulting list of edges is stored to be used for evaluation. In total, there are 4 interventions for all datasets, except for SBM Low Cohesion and SBM High Cohesion. For these two graphs, there are only 3 interventions, due to very low values of polarization and disagreement in the BC-Model with $\epsilon = 0.5$. It was decided to exclude this model from further analysis for these two datasets.

Evaluation of interventions For evaluation, an intervention corresponding to the dataset in focus is loaded and the edges are inserted into the network. Subsequently, all 4 (or 3 in the case of SBM Low Cohesion or SBM High Cohesion) models are run on the new network with the edges of the intervention and, at convergence, polarization and disagreement are measured. Both measurements are compared to the values of the original graph without the intervention. This process is repeated for all datasets in the directed and undirected versions.

And lastly, another algorithm was developed, to the end of constructing mixed interventions to reduce polarization and disagreement across multiple models.

3.4 Algorithms

3.4.1 Greedy algorithm for finding interventions on full search space

Algorithm 3.1: GREEDY(k) on full Search Space

```
$complementaryEdgelist \leftarrow$ List of edges of the full complementary graph of
    G(V,E)
$finalEdges \leftarrow []$
for $i = 0$ to $k - 1$ do
    $pol \leftarrow \sum_{i \in V} \frac{(z_i - \bar{z})^2}{n}$
    for $edge \in complementaryEdgelist$ do
        $E \leftarrow E \cup \{edge\}$
        $z \leftarrow model(G(V, E), s)$
        $polarization \leftarrow \sum_{i \in V} \frac{(z_i - \bar{z})^2}{n}$
        if $polarization < pol$ & $edge \notin finalEdges$ then
            $currentBestEdge \leftarrow edge$
            $pol \leftarrow polarization$
        end
        $E \leftarrow E \setminus \{edge\}$
    end
    $finalEdges \leftarrow finalEdges \cup \{currentBestEdge\}$
end
return $finalEdges$
```

This algorithm was developed to find a set of edges, a so-called intervention, to insert into the graph to reduce polarization in the opinion formation process. Since a node's opinion is influenced by adjacent nodes in both models, adding edges to the graph can heavily influence the opinion formation process and therefore also lead to a higher or lower polarization in the network. So, in adding edges to the network, one can purposefully impact the opinion formation process in different directions. In this thesis, the goal is to reduce the polarization in the opinion profile at convergence. In a real-world setting, reducing polarization could also be the goal of social network administrators or the government.

In this algorithm, the full search space for finding edges to add to the graph is used, which is called the complementary graph $G(V, E)$. The complementary graph of G contains all possible edges, which are not in E. So, depending on the density of the graph, the complementary graph can be quite large, which makes this algorithm computationally very expensive.

As can be seen in the pseudocode of Algorithm 3.1 above, the algorithm first generates a list of edges contained in the complementary graph, which will be searched later. Additionally, an empty list is created as a container for the edges in the intervention. In a loop, which is performed k times, a maximum value of the polarization of the original graph is set for a variable called “pol”. Because the polarization is normalized, it can never be greater than 1.

Afterwards, each edge in the edgelist of the complementary graph is individually inserted into the graph, the model is run with the new edge and the polarization of the resulting opinion profile is calculated. If the polarization is lower than the comparison variable “pol”, the edge is saved as the edge with currently the lowest polarization outcome. If another edge later has an even lower polarization outcome, it will be overwritten and the new edge will be saved. Lastly, the edge is deleted from the graph again.

After all edges in the complementary graph are searched, the smallest edge will be added to the intervention, and this process is performed k times, until k edges have been found to form the intervention to reduce the polarization in the opinion profile. In the end, the intervention is returned.

3.4.2 Greedy algorithm for finding interventions on reduced search space

This algorithm was also developed to find a set of edges, an intervention, to reduce the polarization in the opinion formation process. As can be seen in the section above, the original algorithm on the full search space is, depending on the density of the graph and its complementary graph, computationally very expensive and it takes a long time to run. Therefore, another algorithm on a reduced search space was developed.

The algorithm restricts the search space by generating a list of $10 * k$ candidate edges. To this end, the algorithm generates two lists of the graph’s vertices - one, where the vertices are ordered by the value of their opinion in ascending order and another one where the vertices are ordered by their opinions in descending order.

Then, the candidate edges are collected by connecting one node with a low opinion value with a node with a high opinion value. These edges will potentially lower the polarization in the network, since because of the new edge both nodes will influence each other’s extreme opinions and converge to a more average opinion. An edge is admitted as a candidate edge, if it does not exist in E already.

After the collection of candidate edges is done, the comparison variable “pol” is again set to 1, like in the algorithm above. Each edge is again individually inserted into the graph and the model is run with the new edge to obtain the opinion profile at convergence. With that, the polarization is calculated and if it is smaller than the current maximum value of “pol”, and the edge has not been added to the intervention, then it is saved as the edge with currently the smallest polarization outcome. Afterwards, the edge is taken out of the graph again. This process is repeated for all candidate edges and the edge with the highest polarization reduction is added to the intervention. Lastly, the intervention is returned.

Algorithm 3.2: GREEDY(k) on reduced Search Space

```
$k \in \mathbb{Z}^+$
$candidates \leftarrow []$
$finalEdges \leftarrow []$
$L \leftarrow$ List of vertices ordered by their expressed opinion value $z$ in ascending order
$H \leftarrow$ List of vertices ordered by their expressed opinion value $z$ in descending order
while $candidates < 10 * k$ do
    for $i = 0$ to $\sqrt{k}$ do
        if $(L[i], H[i]) \notin E$ then
            $candidates \leftarrow candidates \cup \{(L[i], H[i])\};$
        end
    end
end
for $i = 0$ to $k - 1$ do
    $pol \leftarrow \sum_{i \in V} \frac{(z_i - \overline{z})^2}{n}$
    for $edge \in candidates$ do
        $E \leftarrow E \cup \{edge\}$
        $z \leftarrow model(G(V, E), s)$
        $polarization \leftarrow \sum_{i \in V} \frac{(z_i - \overline{z})^2}{n}$
        if $polarization < pol$ & $edge \notin finalEdges$ then
            $currentBestEdge \leftarrow edge$
            $pol \leftarrow polarization$
        end
    end
    $finalEdges \leftarrow finalEdges \cup \{currentBestEdge\}$
end
return $finalEdges$
```

CHAPTER 4

Results

In this section, the results of the experiments explained the previous section will be presented and analyzed. First, the results of the experiments without any intervention will be presented. Subsequently, the results of polarization and disagreement with the interventions calculated for each model as well as the results with mixed interventions will be presented. Moreover, a comparison of the results of the algorithms on the full search space and on the reduced search space will be presented.

4.1 Results of the models without intervention

The models were run until the convergence criterion defined above was met and the polarization as well as the disagreement at that point in time were calculated and will be used for comparing the different models. Additionally, the number of time steps necessary to reach convergence will be given for each model.

In this section, the results for each dataset will be described and presented and the discussion of results can be found in the Discussion section below.

4.1.1 Twitter Large

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.1265	0.0329	0.3954	0.3945	0.3894	0.3690
disagreement	0.0542	0.0056	0.8292	0.8229	0.8097	0.7492
convergence	t=9	t=61	t=30	t=131	t=34	t=77

Table 4.1: Results of Twitter Large without intervention

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.3557	0.3227	0.1955	0.0420
disagreement	0.6486	0.6100	0.1168	0.0073
convergence	t=38	t=51	t=27	t=70

Table 4.2: Results of Twitter Large without intervention

The results of the experiments with the Twitter Large dataset can be seen in Table 4.1 and Table 4.2. The stubbornness parameter c of the FJ-Model was set to 1 for this dataset.

If we compare the polarization of all models, we can see that in the directed case, the FJ-Model has the smallest polarization and the BC-Model with an epsilon of 0.05 has the highest polarization. In the undirected case, again the smallest polarization is achieved by the FJ-Model and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$.

If we compare each model in the directed and the undirected case, we can see that the polarization is in the same order of magnitude for all models except the BC-Model with $\epsilon = 0.75$ and the FJ-Model, where the polarization is higher in the directed case compared to the undirected case.

If we compare the disagreement of all models, we can see that in the directed case, the FJ-Model achieves the smallest disagreement and the BC-Model with $\epsilon = 0.05$ achieves the highest disagreement, whereas in the undirected case, the FJ-Model achieves the smallest disagreement and the BC-Model with $\epsilon = 0.05$ achieves the highest disagreement.

Comparing each model in the directed and undirected case, we can see that for the BC-Model with $\epsilon = 0.05$, $\epsilon = 0.2$ and $\epsilon = 0.5$, the disagreement is in the same order of magnitude. However, for the FJ-Model, the disagreement in the directed case is almost 10 times as high as in the undirected case. Moreover, for the BC-Model with $\epsilon = 0.75$, the disagreement in the directed case is 16 times as high as in the undirected case.

Comparing the time it takes for a model to reach convergence, we can see that for all models, the undirected case takes longer to convergence. Specifically, it can take up to 6 times longer than in the directed case, which can be seen in the FJ-Model.

4.1.2 SBM High Cohesion

The results of the experiments for the SBM High Cohesion dataset can be found in Tables 4.3 and 4.4. When avg_deg_G is the average degree of a graph G, the stubbornness parameter c of the FJ-Model was set to $c = 2 \cdot avg\_deg_{SBM\,High\,Cohesion}$ for this dataset.

What is interesting about this dataset, is that for the BC-Model with $\epsilon = 0.75$, all values of polarization and disagreement in both the directed and the undirected case are 0.

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.1808	0.1937	0.2076	0.2070	0.1856	0.1858
disagreement	0.0537	0.0638	0.0766	0.0764	0.0379	0.0379
convergence	t=3	t=3	t=10	t=8	t=5	t=5

Table 4.3: Results of SBM High Cohesion without intervention and initialization of opinions per communities

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.1725	0.1708	0.0000	0.0000
disagreement	0.0352	0.0349	0.0000	0.0000
convergence	t=21	t=20	t=13	t=13

Table 4.4: Results of SBM High Cohesion without intervention and initialization of opinions per communities

Therefore, this model will be excluded from further analysis, as it cannot be used for the purpose of reducing polarization and disagreement.

Comparing the polarization of all remaining models, in the directed case, the FJ-Model achieves the smallest polarization and the BC-Model with $\epsilon = 0.05$ achieves the highest polarization, whereas in the undirected case, the smallest polarization is achieved by the FJ-Model and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$.

Comparing the directed and the undirected case for each model, we can see that for the BC-Models, the polarization is in the same order of magnitude for both cases, whereas in the FJ-Model, the polarization is 3 times higher in the directed case compared to the undirected case.

Furthermore, if we compare the disagreement of all models, in the directed case, the smallest disagreement is achieved by the FJ-Model and the highest disagreement is achieved by the BC-Model with $\epsilon = 0.05$. In the undirected case, again the FJ-Model achieves the smallest disagreement and the BC-Model with $\epsilon = 0.05$ achieves the highest disagreement.

If we compare the directed and the undirected case for each model, we can see that again for the BC-Models the disagreement is in the same order of magnitude for both cases, but for the FJ-Model, the disagreement is 4 times higher in the directed case compared to the undirected case.

Regarding convergence, we can see that all models converge at more or less the same time in both the directed and the undirected case.

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.1463	0.2085	0.3204	0.3213	0.2754	0.2765
disagreement	0.2848	0.4118	0.6454	0.6477	0.5501	0.5523
convergence	t=3	t=3	t=14	t=11	t=14	t=18

Table 4.5: Results for SBM High Cohesion with random opinion initialization

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.0000	0.0000	0.0000	0.0000
disagreement	0.0000	0.0000	0.0000	0.0000
convergence	t=10	t=10	t=8	t=8

Table 4.6: Results for SBM High Cohesion with random opinion initialization

The results for the SBM High Cohesion dataset with initialization of the opinions by drawing from a uniform distribution can be found in Tables 4.5 and 4.6. In contrast to the results with the same model presented above, we can see that now for both BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$ all values of polarization and disagreement in both the directed and the undirected case are 0, which means that they will be excluded from further analysis, as they cannot be used for the purpose of reducing polarization and disagreement.

Again, the FJ-Model achieves the lowest values in both polarization and disagreement across all models, and the BC-Model with $\epsilon = 0.05$ achieves the highest values of polarization and disagreement across all models. With this configuration, the values of polarization and disagreement are roughly in the same order of magnitude for both the directed and undirected case across all models.

For further experiments, only the initialization of opinions with differences per communities will be considered.

4.1.3 SBM Low Cohesion

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.1717	0.1919	0.2172	0.2172	0.1993	0.1990
disagreement	0.0933	0.1097	0.1297	0.1293	0.0929	0.0928
convergence	t=20	t=22	t=21	t=15	t=7	t=6

Table 4.7: Results for SBM Low Cohesion with initialization of opinions per community

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.0000	0.0000	0.0000	0.0000
disagreement	0.0000	0.0000	0.0000	0.0000
convergence	t=48	t=42	t=29	t=29

Table 4.8: Results for SBM Low Cohesion with initialization of opinions per community

The results of the experiments with the SBM Low Cohesion dataset can be found in Tables 4.7 and 4.8. When avg_deg_G is the average degree of a graph G, the stubbornness parameter c of the FJ-Model was set to $c = 2 \cdot avg\_deg_{SBM\,Low\,Cohesion}$ for this dataset.

As we can see in Table 4.8, for both BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$, all values for polarization and disagreement in the directed and the undirected case are 0, which is why these models will be excluded from further analysis, since they do not serve the purpose of reducing polarization and disagreement.

Comparing the polarization across all models, we can see that in the directed case, the smallest polarization is achieved by the FJ-Model and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$. Similarly, in the undirected case, the smallest polarization is achieved by the FJ-Model and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$.

If we compare the directed and undirected case for each model, we can see that the polarization is in the same order of magnitude for both remaining BC-Models, but in the FJ-Model, the polarization is 3 times higher in the directed case compared to the undirected case.

Moreover, comparing the disagreement across all models, in the undirected case, the smallest disagreement is achieved by the FJ-Model and the highest disagreement is achieved by the BC-Model with $\epsilon = 0.05$. In the undirected case, the smallest disagreement is achieved by the FJ-Model and the highest disagreement is achieved by the BC-Model with $\epsilon = 0.05$.

Considering convergence, we can see that convergence time is in the same order of magnitude across all models in both the directed and the undirected case.

If we look at Tables 4.9 and 4.10, we find the results of the experiments with the SBM Low Cohesion dataset, but with initialization of the opinions by drawing from a uniform distribution. Again, we can see that for the BC-Model with $\epsilon = 0.75$, all values for polarization and disagreement in both the directed and the undirected case are 0, which is why this model will be excluded from further analysis, as it cannot be used for the purpose of reducing polarization and disagreement. For similar reasons, the BC-Model with $\epsilon = 0.5$ will be excluded from further analysis, since polarization and disagreement are 0 in the undirected case and very close to 0 in the directed case.

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.1550	0.2193	0.3372	0.3405	0.2732	0.2670
disagreement	0.3031	0.4365	0.6897	0.6952	0.5450	0.5338
convergence	t=10	t=10	t=31	t=34	t=24	t=25

Table 4.9: Results for SBM Low Cohesion with random opinion initialization

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.0018	0.0000	0.0000	0.0000
disagreement	0.0020	0.0000	0.0000	0.0000
convergence	t=19	t=20	t=12	t=12

Table 4.10: Results for SBM Low Cohesion with random opinion initialization

Similarly to the above results, the FJ model achieves the smallest values of polarization and disagreement across all models, while the BC model with $\epsilon = 0.05$ produces the highest values of polarization and disagreement across all remaining models.

Furthermore, across all models, both polarization and disagreement are roughly in the same order of magnitude in both the directed and the undirected case. For further experiments, only the initialization of opinions with differences per communities will be considered.

4.1.4 Reddit Data

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.0002	0.0000	0.0009	0.0006	0.0004	0.0002
disagreement	0.0003	0.0000	0.0020	0.0015	0.0008	0.0005
convergence	t=10	t=4	t=7	t=6	t=8	t=4

Table 4.11: Results of Reddit data without intervention

The results of the experiments with the Reddit data can be found in Tables 4.11 and 4.12. The stubbornness parameter c of the FJ Model was set to 1 for this dataset.

Looking at the values of polarization and disagreement across all models, we can see that the BC-Model with $\epsilon = 0.75$ produces 0 for polarization and disagreement in the undirected case, and in the directed case, the polarization is 0.0002 and the disagreement 0.0003. This model will be excluded from further analysis since it is not suitable for investigating reducing polarization and disagreement.

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.0002	0.0000	0.0002	0.0000
disagreement	0.0003	0.0000	0.0003	0.0000
convergence	t=8	t=4	t=8	t=4

Table 4.12: Results of Reddit data without intervention

If we compare the polarization across all models, in the directed case, the FJ-Model achieves the smallest polarization and the BC-Model with $\epsilon = 0.05$ achieves the highest polarization. In the undirected case, the FJ-Model achieves the smallest polarization, along with the BC-Model with $\epsilon = 0.5$ and the BC-Model with $\epsilon = 0.05$ achieves the highest polarization.

Generally, we can see a trend that the models converge faster in the undirected case on this specific dataset than in the directed case. Regarding the differences between the directed and undirected cases for each model, we can see that the polarization is up to 10 times higher in the directed case compared to the undirected case.

If we compare the disagreement across all models, we can see that in the directed case, the FJ-Model achieves the smallest disagreement along with the BC-Model with $\epsilon = 0.5$ and the BC-Model with $\epsilon = 0.05$ achieves the highest disagreement, and in the undirected case, both the FJ-Model and the BC-Model with $\epsilon = 0.5$ achieve the smallest disagreement, and the BC-Model with $\epsilon = 0.05$ achieves the highest disagreement.

This dataset will be excluded from further analysis altogether due to the very small values of polarization and disagreement across all models.

4.1.5 Flixster data

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.0002	0.0593	0.0129	0.3310	0.0154	0.0015
disagreement	0.0004	0.0225	0.3875	0.6418	0.0000	0.0000
convergence	t=6	t=32	t=32	t=1338	t=61	t=419

Table 4.13: Results for Flixster data without intervention

In Tables 4.13 and 4.14, the results for the Flixster social network data can be found. The stubbornness parameter c of the FJ-Model was set to 1 for this dataset.

If we compare the polarization across all models, we can see that in the directed case, the BC-Model with $\epsilon = 0.75$ achieves the smallest polarization and the BC-Model with $\epsilon = 0.2$ achieves the highest polarization, while in the undirected case, the smallest

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.0056	0.0013	0.0000	0.0013
disagreement	0.0000	0.0000	0.0000	0.0000
convergence	t=67	t=200	t=26	t=221

Table 4.14: Results for Flixster data without intervention

polarization is achieved by the BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$ and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$.

Interestingly, as opposed to the other datasets, with this dataset we notice considerable differences between the directed and undirected cases in some of the models. For example, in the FJ-Model, the polarization is in the undirected case 296 times higher than in the directed case. However, in the BC-Model with $\epsilon = 0.05$, the polarization is 25 times higher in the undirected case than in the directed case and in the BC-Model with $\epsilon = 0.2$, polarization is 10 times higher in the directed case compared to the undirected case. In the BC-Model with $\epsilon = 0.5$, the polarization is 4 times higher in the directed case than in the undirected case, while in the BC-Model with $\epsilon = 0.75$, the polarization in the directed case is 0, but not in the undirected case.

If we compare the disagreement across all models, we can see that, regardless of directed or undirected cases, the value of disagreement becomes 0 for the BC-Models with $\epsilon = 0.2$, $\epsilon = 0.5$ and $\epsilon = 0.75$. Consequently, these models have the smallest value of disagreement, but the BC-Model with $\epsilon = 0.05$ achieves the highest disagreement in both the directed and undirected case.

Regarding the differences in disagreement between the directed and undirected case, we can see that there is no difference for the BC-Models with $\epsilon = 0.2$, $\epsilon = 0.5$ and $\epsilon = 0.75$, since all disagreement values are 0. However, in the BC-Model with $\epsilon = 0.05$, disagreement is about twice as high in the undirected case compared to the directed case. In the FJ-Model, disagreement is 56 times higher in the undirected case compared to the directed case.

Considering convergence, we can see that especially the BC-Model with $\epsilon = 0.05$ took a very long time, specifically 1338 steps, to converge. We can see a trend that the model takes longer to converge in the undirected case compared to the directed case. Moreover, convergence times are faster for the FJ-Model than the BC-Models.

This dataset will be excluded from further analysis altogether due to time constraints and limited computational resources.

4.1.6 Barabasi-Albert-Graph ($m = 10$)

In Tables 4.15 and 4.16, the results of the experiments with the Barabasi-Albert-Graph with $m = 10$ can be found. When avg_deg_G is the average degree of a graph G, the

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.2141	0.2180	0.3255	0.3252	0.3117	0.2690
disagreement	0.3056	0.3724	0.6707	0.6708	0.6135	0.5332
convergence	t=7	t=25	t=10	t=13	t=13	t=37

Table 4.15: Results for Barabasi-Albert-Graph with m=10 without intervention

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.2477	0.0000	0.1533	0.0000
disagreement	0.3921	0.0000	0.1622	0.0000
convergence	t=34	t=12	t=18	t=8

Table 4.16: Results for Barabasi-Albert-Graph with $m = 10$ without intervention

stubbornness parameter c of the FJ-Model was set to $c = 2 \cdot avg\_deg_{BAG10}$ for this dataset.

If we compare the polarization across all models, we can see that in the directed case, the BC-Model with $\epsilon = 0.75$ achieves the smallest polarization and the BC-Model with $\epsilon = 0.05$ achieves the highest polarization, whereas in the undirected case, the smallest polarization is achieved by the BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$, and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$.

Comparing the results of polarization and disagreement in both the directed and undirected case across all models, we can see that for the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$ and the FJ-Model, polarization and disagreement are roughly in the same order of magnitude, while for the other two BC-Models, both measurements are 0 in the undirected case, but not in the directed case.

Regarding the disagreement across all models, we can see that in the directed case, the smallest disagreement is achieved by the BC-Model with $\epsilon = 0.75$ and the highest disagreement is achieved by the BC-Model with $\epsilon = 0.05$. In the undirected case, the BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$ achieve the smallest disagreement and the BC-Model with $\epsilon = 0.05$ achieves the highest disagreement.

Considering convergence, the results are rather ambiguous across all models, since for some models, the directed case converges faster and in others, the undirected case converges faster. Hence, there is no general trend that is recognizable.

4.1.7 Barabasi-Albert-Graph ($m = 20$)

In Tables 4.17 and 4.18, the results of the experiments with Barabasi-Albert-Graph with $m = 20$ without intervention can be found. When avg_deg_G is the average degree of a

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.2048	0.2155	0.3252	0.3233	0.3093	0.2539
disagreement	0.3082	0.3841	0.6678	0.6645	0.5874	0.5192
convergence	t=8	t=16	t=11	t=32	t=22	t=27

Table 4.17: Results for Barabasi-Albert-Graph with m=20 without intervention

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.1592	0.0000	0.0987	0.0000
disagreement	0.1981	0.0000	0.1042	0.0000
convergence	t=22	t=37	t=17	t=7

Table 4.18: Results for Barabasi-Albert-Graph with m=20 without intervention

graph G, the stubbornness parameter c of the FJ-Model was set to $c = 2 \cdot avg\_deg_{BAG\,20}$ for this dataset.

Comparing the polarization across all models, we can see that in the directed case, the smallest polarization is achieved by the BC-Model with $\epsilon = 0.75$ and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$. In the undirected case, the smallest polarization is achieved by the BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$ and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$.

If we compare the polarization and disagreement in the directed and undirected case across all models, we can see that for the BC-Model with $\epsilon = 0.05$ and $\epsilon = 0.2$ and the FJ-Model both measurements are in the same order of magnitude, while for the other BC-Models, both measurements are 0 in the undirected case, but not in the directed case.

Regarding disagreement across all models, it can be seen that in the directed case, the smallest disagreement is achieved by the BC-Model with $\epsilon = 0.75$ and the highest disagreement is achieved by the BC-Model with $\epsilon = 0.05$, while in the undirected case, the smallest disagreement is achieved by the BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$ and the highest disagreement is achieved by the BC-Model with $\epsilon = 0.05$.

If we look at convergence times, we can see that the models tend to converge faster in the directed case, except the BC-Model with $\epsilon = 0.75$, which converges faster in the undirected case.

4.1.8 Barabasi-Albert-Graph ($m = 30$)

The results of the experiments with the Barabasi-Albert-Graph with $m = 30$ without intervention can be found in Tables 4.19 and 4.20. When avg_deg_G is the average degree of

	FJ-Model		BC-Model ($\epsilon = 0.05$)		BC-Model ($\epsilon = 0.2$)	
	directed	undirected	directed	undirected	directed	undirected
polarization	0.1979	0.2151	0.3282	0.3234	0.2941	0.2704
disagreement	0.3026	0.3875	0.6635	0.6541	0.5651	0.5385
convergence	t=7	t=5	t=15	t=44	t=28	t=14

Table 4.19: Results for Barabasi-Albert-Graph with $m = 30$ without intervention

	BC-Model ($\epsilon = 0.5$)		BC-Model ($\epsilon = 0.75$)	
	directed	undirected	directed	undirected
polarization	0.2183	0.0000	0.0767	0.0000
disagreement	0.3954	0.0000	0.0836	0.0000
convergence	t=18	t=38	t=18	t=7

Table 4.20: Results for Barabasi-Albert-Graph with $m = 30$ without intervention

a graph G, the stubbornness parameter c of the FJ-Model was set to $c = 2 \cdot avg\_deg_{BAG30}$ for this dataset.

Comparing the polarization across all models, we can see that in the directed case, the BC-Model with $\epsilon = 0.75$ achieves the smallest polarization and the BC-Model with $\epsilon = 0.05$ achieves the highest polarization. In the undirected case, the smallest polarization is achieved by the BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$ and the highest polarization is achieved by the BC-Model with $\epsilon = 0.05$.

If we compare the values of polarization and disagreement in the directed and undirected case across all models, we can see that again for the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$ and the FJ-Model, polarization and disagreement are in the same order of magnitude for both the directed and undirected case. Also, in the BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$, both polarization and disagreement are 0 in the undirected case, but not in the directed case.

Regarding disagreement, we can see that in the directed case, the smallest value of disagreement is achieved by the BC-Model with $\epsilon = 0.75$ and the highest disagreement is achieved by the BC-Model with $\epsilon = 0.05$. In the undirected case, the smallest disagreement is achieved by the BC-Models with $\epsilon = 0.5$ and $\epsilon = 0.75$ and the highest disagreement is achieved by the BC-Model with $\epsilon = 0.05$.

Considering convergence times, there is no clear trend as to which model converges faster in which case - directed or undirected. While the FJ-Model, the BC-Model with $\epsilon = 0.2$ and the BC-Model with $\epsilon = 0.75$, converge faster in the directed case, the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.5$, converge faster in the undirected case. This could be due to other unknown factors contributing towards convergence times.

4.2 Comparison of Interventions

In this section, the results of comparing interventions across different models will be presented and described, while the discussion of results can be found in the Discussion section below.

In the following tables, the relative change of polarization and disagreement after inserting a specific intervention into the graph is reported. Consider the polarization p, calculated as defined above, of the opinion profile without any intervention and the polarization p' of the opinion profile after the intervention was inserted into the graph. The relative change in polarization is defined as the ratio $\frac{p'}{p}$.

Analogously, consider the disagreement d, calculated as defined above, of the opinion profile without any intervention and the disagreement d' of the opinion profile after the intervention was inserted into the graph. The relative change in disagreement is then defined as the ratio $\frac{d'}{d}$.

The interventions for the datasets BAG 10, BAG 20 and BAG 30 were computed using the greedy algorithm described above on the full search space with $k = 50$ for the BC-Model with $\epsilon = 0.05$ and $k = 20$ for all other models. The interventions for the SBM Low Cohesion and SBM High Cohesion datasets were computed using the greedy algorithm described above on the full search space with $k = 20$. The interventions for the Twitter Large dataset was, due to time and computational constraints, computed using the greedy algorithm on the reduced search space using $k = 100$ in the FJ-Model and $k = 20$ in the BC-Models.

Each column represents one intervention that has been calculated with the model specified at the top of the column. Each row in that column represents the relative change of polarization and disagreement after inserting said column into the graph in a specific model.

For each dataset, there will be four different tables: one table each for polarization in the directed and undirected cases of the graph, and one table each for polarization in the directed and undirected cases of the graph.

4.2.1 Twitter Large

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	**0.992134**	0.998816	0.997658	0.996601
	BC 0.05	1.000000	**0.999413**	1.000000	1.000000
	BC 0.2	1.000000	0.986563	**0.986501**	0.999885
	BC 0.5	1.000000	0.997720	0.996698	**0.993962**

Table 4.21: Reduction of polarization in Twitter Large directed

In Table 4.21, the relative change in polarization can be seen in the Twitter Large data set after inserting various interventions. We can see that for each model (each row), the intervention that was optimized for the model achieves the greatest reduction in polarization. For example, if we look at the row with the index “BC 0.05”, we can see that the intervention calculated with the FJ-Model does not reduce the polarization measure at all in the BC-Model with $\epsilon = 0.05$, while the intervention calculated with the BC-Model with $\epsilon = 0.05$ reduces the polarization by around 0.1%. What is more, both interventions calculated with the BC-Models with $\epsilon = 0.2$ and $\epsilon = 0.5$ achieve no reduction in polarization.

Additionally, we can see that inserting an intervention calculated with one model does not increase polarization in a different model, in fact, sometimes there is no change at all. For example, if we look at the intervention calculated with the FJ-Model (the column named “FJ”), we can see that there is no change in polarization at all in the bounded-confidence models. Again in Table 4.22, we can see that the interventions calculated with a specific model achieve the greatest reduction in polarization in said model, while there is no change or a very small reduction in the other models. Concerning the development of disagreement after inserting an intervention, we can see that in Table 4.23 in the directed case, for example in the FJ-Model, disagreement declines as well as polarization. While this is not typically the case, in this case the reduction is marginal. Also in the undirected case, in Table 4.24 we can see that disagreement decreases slightly in the FJ-Model, but it increases for example if we insert the intervention calculated with the BC-Model with $\epsilon = 0.05$ into other bounded-confidence models with different values for the confidence bound parameter.

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	**0.964339**	0.999408	0.998991	0.998780
	BC 0.05	1.000000	**0.999885**	1.000000	1.000000
	BC 0.2	1.000000	0.999580	**0.998918**	1.000000
	BC 0.5	1.000000	0.999635	0.999481	**0.998793**

Table 4.22: Reduction of polarization in Twitter Large undirected

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	0.996458	0.990032	0.988503	0.991198
	BC 0.05	1.000903	0.998506	0.999928	0.999944
	BC 0.2	1.000861	0.967645	0.968425	0.999807
	BC 0.5	1.000735	0.996154	0.995124	0.991243

Table 4.23: Disagreement with intervention in Twitter Large directed

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	0.994178	0.999884	0.999896	0.999815
	BC 0.05	1.001230	1.000051	0.999928	0.999944
	BC 0.2	1.000979	1.000118	1.000019	0.999938
	BC 0.5	1.000818	1.000093	1.000076	0.999895

Table 4.24: Disagreement with intervention Twitter Large undirected

4.2.2 BAG 10

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	**0.975654**	0.996135	0.996485	0.996725
	BC 0.05	1.000000	**0.965089**	0.998981	1.000017
	BC 0.2	1.000000	0.923332	**0.884584**	1.002303
	BC 0.5	1.000000	0.906533	0.835240	**0.612760**

Table 4.25: Reduction of polarization in BAG 10 directed

In Table 4.25 the relative change in polarization in the directed Barabasi-Albert-Graph with parameter $m = 10$ after inserting different interventions can be seen. Like in the example with Twitter Large above, we can see that for each model, the intervention calculated with that specific model yields the biggest reduction in polarization, while in other models, the polarization is not changed at all or there is even a slight increase. In Table 4.26 we can see the relative change in polarization in the undirected case of the Barabasi-Albert graph described above. Compared to the directed case, we can see that the reduction of polarization in the bounded-confidence model with $\epsilon = 0.5$ is even greater in the undirected case, however, this result should be taken with caution, since the baseline polarization for this graph with this model is 0.0000 (rounded). This is a very low value for polarization, since it is always in $[0, 1]$ through normalization and the relative reduction in polarization might appear large (65%), but in absolute values, it is

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	**0.989020**	0.999652	0.999934	0.999748
	BC 0.05	1.000000	**0.996331**	0.999996	1.000000
	BC 0.2	1.000000	1.007117	**0.990924**	1.000146
	BC 0.5	0.868289	1.425110	21768.382005	**0.353043**

Table 4.26: Reduction of polarization in BAG 10 undirected

not. This also explains the extremely large value in the BC-Model with $\epsilon = 0.5$ with the intervention calculated with the BC-Model with $\epsilon = 0.2$. We can also see that, overall, the reductions of polarization are larger in the directed graph than in the undirected graph.

In Table 4.27 we can see the relative change in disagreement in the directed Barabasi-Albert graph with $m = 10$ after inserting different interventions. It can be seen that the intervention calculated with the FJ-Model creates an increase in disagreement in all models (first column). This behavior is expected, since the intervention reduces polarization by inserting edges between nodes with very different initial opinions, which simultaneously leads to a higher disagreement. And finally, Table 4.28 shows the relative change in disagreement in the undirected case of BAG 10 after inserting different interventions. In the FJ-Model, we can see that the intervention calculated with the same model yields an increase in disagreement, which is expected for the reasons explained above, while the other interventions cause a slight decrease in disagreement.

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	1.020295	0.989698	0.994810	0.995331
	BC 0.05	1.018249	0.944693	0.994697	0.996909
	BC 0.2	1.020331	0.898347	0.855533	1.008919
	BC 0.5	1.029548	0.938207	0.761066	0.465037

Table 4.27: Disagreement with intervention in BAG 10 directed

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	1.017301	0.990940	0.996102	0.996286
	BC 0.05	1.018358	0.991712	0.996061	0.996863
	BC 0.2	1.011842	0.998795	0.987581	0.998129
	BC 0.5	0.881966	1.465966	23547.212643	0.352772

Table 4.28: Disagreement with intervention on BAG 10 undirected

4.2.3 BAG 20

In Table 4.29 the relative change in polarization in BAG 20 is shown after inserting different interventions. Again, it can be seen that for each model, the intervention calculated with that specific model achieves the greatest reduction in polarization, while other interventions yield either a smaller reduction or no reduction at all.

The greatest reduction is achieved by the bounded-confidence model with $\epsilon = 0.5$, which can be explained with the same reasoning given above.

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	**0.985192**	0.998201	0.998345	0.999267
	BC 0.05	1.000000	**0.938319**	0.998605	0.999625
	BC 0.2	1.000000	0.867608	**0.771384**	0.985186
	BC 0.5	1.000000	0.765676	0.781306	**0.519092**

Table 4.29: Reduction of polarization in BAG 20 directed

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	**0.994077**	0.999935	0.999950	0.999835
	BC 0.05	1.000000	**0.993451**	1.000000	1.000000
	BC 0.2	1.000000	1.004690	**0.982712**	0.999511
	BC 0.5	0.970393	1.107683	0.917036	**0.355845**

Table 4.30: Reduction of polarization in BAG 20 undirected

In Table 4.30, the relative change of polarization in the undirected version of BAG 20 after inserting different interventions is displayed. Compared to the table above displaying the directed case, the greatest reduction is also achieved by the bounded-confidence model with $\epsilon = 0.5$, but the reduction is even greater in the undirected case than in the directed case. Again, this phenomenon can be explained with the same reasoning given above.

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	1.009805	0.994817	0.997043	0.997383
	BC 0.05	1.009388	0.979058	0.997525	0.997505
	BC 0.2	1.010957	0.946683	0.981953	1.005678
	BC 0.5	1.032718	0.825494	0.864220	0.879469

Table 4.31: Disagreement with intervention BAG 20 directed

Tables 4.31 and 4.32 show the relative change in disagreement in the directed and undirected cases of BAG 20 after inserting different interventions. Again, we can see that inserting the interventions causes a slight increase in disagreement in some models, for example inserting the intervention calculated with the BC-Model with $\epsilon = 0.05$ in the FJ-Model. In other models, inserting the interventions leads to a negligible decrease in disagreement.

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	1.009109	0.995005	0.998028	0.998179
	BC 0.05	1.008978	0.994329	0.998041	0.998463
	BC 0.2	1.003450	0.998521	0.983669	0.999907
	BC 0.5	0.975145	1.106376	0.917717	0.361009

Table 4.32: Disagreement with intervention BAG 20 undirected

4.2.4 BAG 30

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	**0.988927**	0.997312	0.998875	0.999119
	BC 0.05	1.000000	**0.962509**	0.998741	0.999252
	BC 0.2	1.000000	0.900483	**0.729284**	1.065636
	BC 0.5	1.000000	0.847582	0.407256	**0.269006**

Table 4.33: Reduction of polarization in BAG 30 directed

In Table 4.33 the relative change in polarization in the directed Barabasi-Albert graph with $m = 30$ after inserting different interventions is displayed. For each model, the greatest reduction is achieved by the intervention calculated with that very model. The greatest reduction overall of about 74% is achieved by the bounded-confidence model with $\epsilon = 0.5$, which is due to the reasons explained above. In Table 4.34, the relative change in polarization after inserting different interventions into the undirected version of the graph mentioned above is displayed. Compared to the directed graph, we can see that the reductions in the BC-Models with $\epsilon = 0.2$ and $\epsilon = 0.5$ are smaller in the undirected graph.

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	**0.995965**	0.999926	0.999971	0.999898
	BC 0.05	1.000000	**0.994667**	0.999445	1.000000
	BC 0.2	1.000000	0.994643	**0.986640**	0.996196
	BC 0.5	0.988540	1.085311	1.097683	**0.561943**

Table 4.34: Reduction of polarization in BAG 30 undirected

Again, we can see that for each model, the intervention calculated with that very model achieves the greatest reduction in polarization, while other interventions either create a smaller reduction, a slight increase in polarization or no change at all. The

BC-Model with $\epsilon = 0.5$ achieves the largest reduction of polarization out of all the models.

In Tables 4.35 and 4.36 the relative change in disagreement after inserting different interventions into the directed and undirected Barabasi-Albert graphs with $m = 30$ is shown.

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	1.006149	0.995567	0.998193	0.998398
	BC 0.05	1.006638	0.953898	0.996624	0.998231
	BC 0.2	1.007862	0.891406	0.666674	1.082631
	BC 0.5	1.009699	0.826067	0.261325	0.167861

Table 4.35: Disagreement with intervention in BAG 30 directed

In the directed case, inserting the intervention calculated with the FJ-Model leads to an increase of disagreement in all models, while in the undirected case, it leads to an increase in disagreement in all models but one, namely the BC-Model with $\epsilon = 0.5$.

Interestingly, in the BC-Model with $\epsilon = 0.5$, inserting the intervention calculated with that same model leads to a decrease in both polarization and disagreement, as already discussed above. This behavior is not expected, since with decreasing polarization an increase in disagreement is expected, though it could be explained through the reasoning about small polarization values given above.

		Interventions			
		FJ	BC 0.05	BC 0.2	BC 0.5
Models	FJ	1.006092	0.996676	0.998651	0.998751
	BC 0.05	1.005953	0.997786	0.997121	0.998898
	BC 0.2	1.003651	0.992635	0.987309	0.993404
	BC 0.5	0.988277	1.086697	1.094780	0.572783

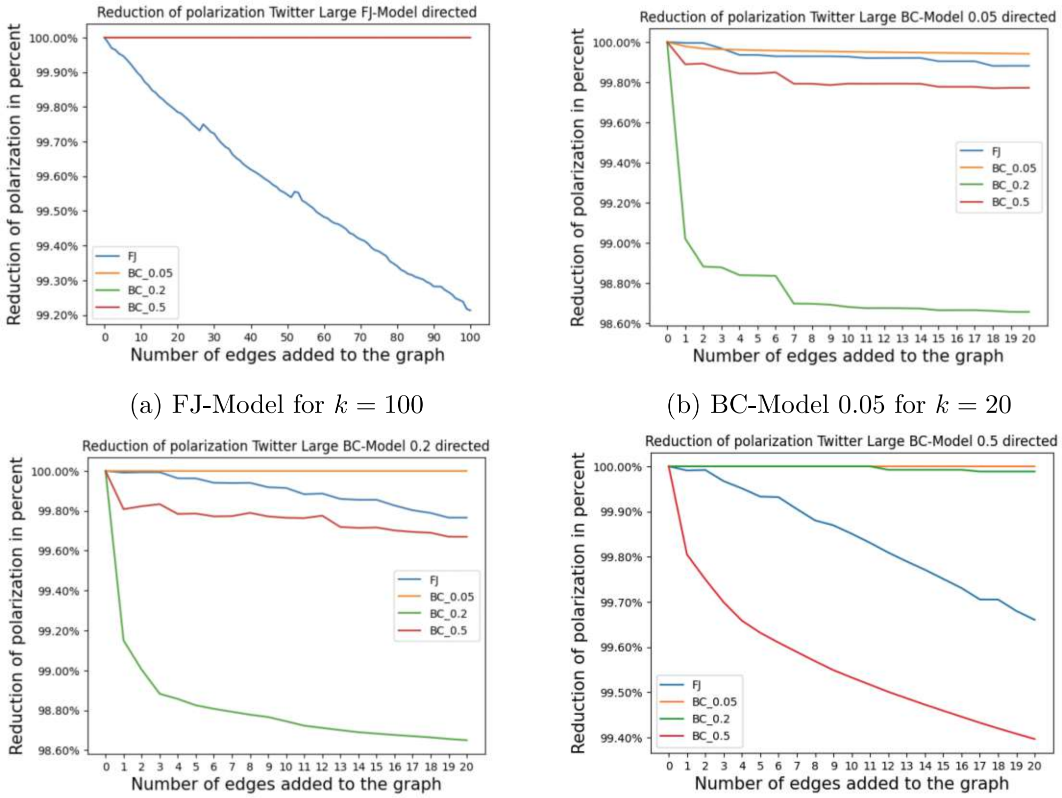
Table 4.36: Disagreement with intervention in BAG 30 undirected

There will be no results reported on the comparison of interventions for the two datasets SBM Low Cohesion and SBM High Cohesion, because the reduction in polarization is very small and therefore the results are not interesting to report.

4.3 Comparison of models

In this section, the results of adding the same intervention to different models with different parameter settings will be presented.

4.3.1 Twitter Large



(a) FJ-Model for $k = 100$

(b) BC-Model 0.05 for $k = 20$

(c) BC-Model 0.2 for $k = 20$

(d) BC-Model 0.5 for $k = 20$

Figure 4.1: Interventions in Twitter Large directed calculated with different models

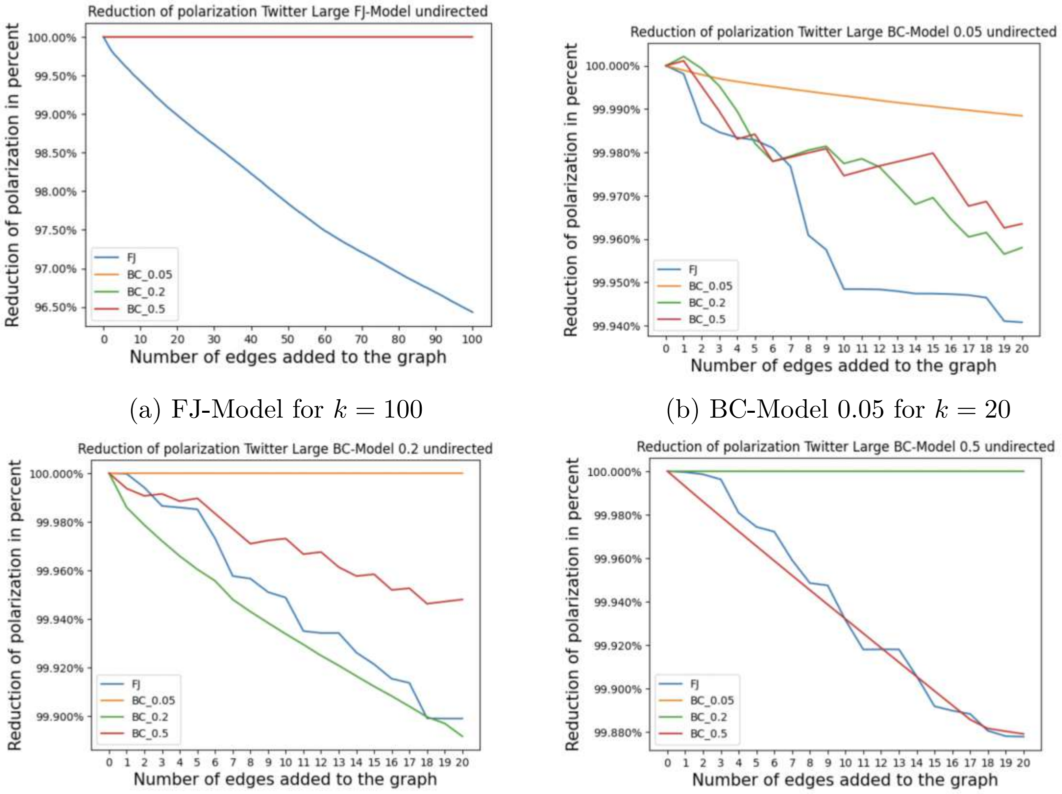
In Figure 4.1 the reduction of polarization in percent depending on the number of edges added to the graph on the directed Twitter Large graph can be seen. Each picture displays the reduction of polarization in each model after inserting the intervention calculated with the model specified beneath the picture into the graph. The interventions for Twitter Large were obtained with the greedy algorithm specified above on the reduced search space, with $k = 100$ for the FJ-Model and $k = 20$ for the other models.

For example, in Figure 4.1a, we can see that the intervention calculated with the FJ-Model reduces polarization only in the FJ-Model, while there is no reduction or increase visible in the other models. This is an interesting result, as we can see that even by inserting interventions calculated with other models there is no "damage" done, i.e. an increase in polarization, but rather no change at all.

Furthermore, we can see that in Figure 4.1b the bounded-confidence model with $\epsilon = 0.2$ achieves the largest reduction in polarization, although the intervention is optimized for the bounded-confidence model with $\epsilon = 0.05$. In Figure 4.1c, the BC-Model with

$\epsilon = 0.2$ achieves the largest reduction, which the intervention is optimized for, while the reduction in the other models is very small, or in the case of the BC-Model with $\epsilon = 0.05$, none at all.

Lastly, the intervention calculated with the BC-Model with $\epsilon = 0.5$ achieves the largest reduction in polarization in that same model, but also a smaller reduction in the FJ-Model, while there is almost no reduction or increase at all in the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$.



(a) FJ-Model for $k = 100$

(b) BC-Model 0.05 for $k = 20$

(c) BC-Model 0.2 for $k = 20$

(d) BC-Model 0.5 for $k = 20$

Figure 4.2: Interventions in Twitter Large undirected calculated with different models

In Figure 4.2, the reduction of polarization in percent depending on the number of edges added to the graph for different interventions in the undirected Twitter Large graph can be seen.

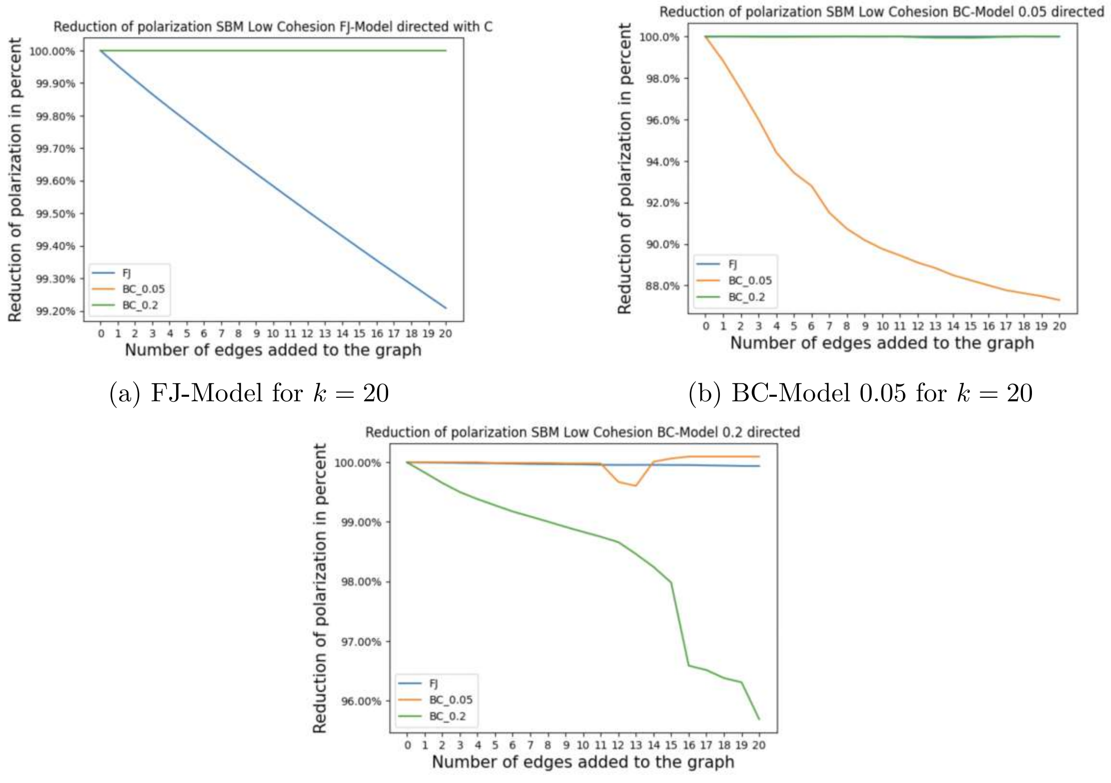
The first intervention to be analyzed was calculated with the FJ-Model as depicted in Figure 4.2a. This intervention reduces polarization only in the FJ-Model, while there is neither a reduction nor an increase visible in the other models.

The experiments with the intervention calculated with the BC-Model with $\epsilon = 0.05$ are analyzed in Figure 4.2b. While there is a small reduction in the model the intervention was optimized for, all three other models achieve a greater reduction.

The experiments with the intervention calculated with the BC-Model with $\epsilon = 0.2$ are depicted in Figure 4.2c. This intervention achieves the greatest reduction with the model it was optimized for, whereas there is a small reduction in the FJ-Model and the BC-Model with $\epsilon = 0.5$. In the BC-Model with $\epsilon = 0.05$, there is neither a reduction nor an increase in polarization.

In Figure 4.2d the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.5$ are depicted. The greatest reduction is achieved by the model this intervention was optimized for. While in the FJ-Model and in the BC-Model with $\epsilon = 0.2$ there is also a small reduction visible, there is neither a reduction nor an increase visible in the BC-Model with $\epsilon = 0.05$.

4.3.2 SBM Low Cohesion



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 20$

(c) BC-Model 0.2 for $k = 20$

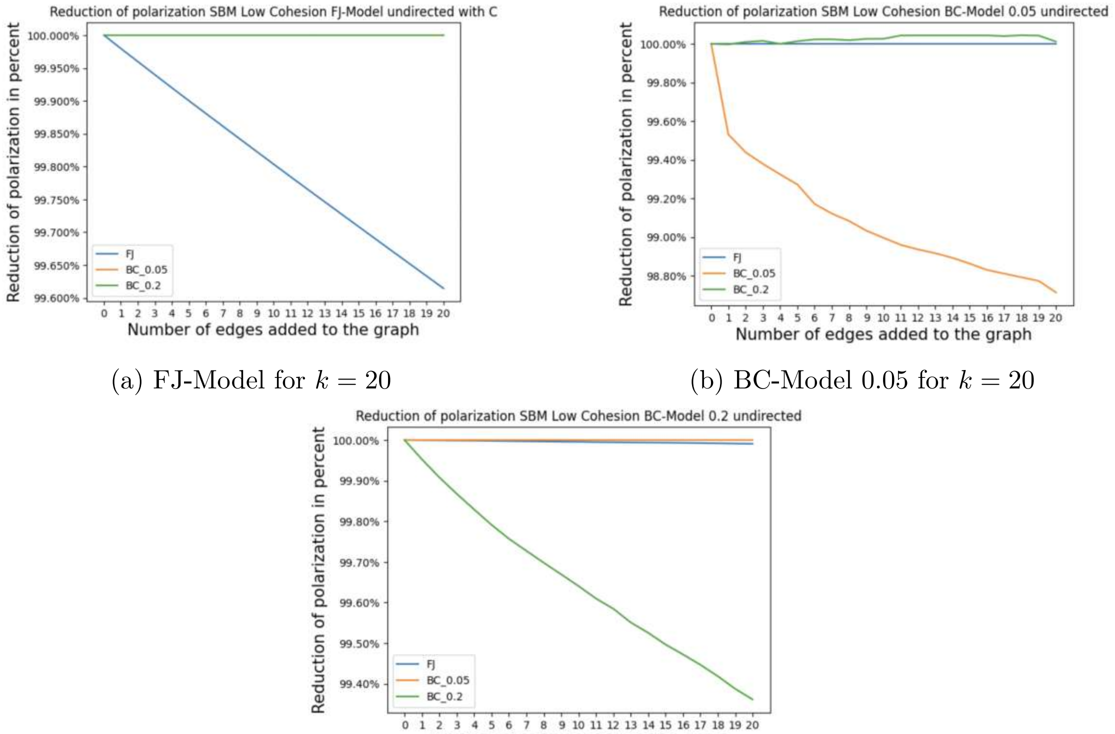
Figure 4.3: Interventions in SBM Low Cohesion directed calculated with different models

In Figure 4.3, the reduction of polarization in percent depending on the number of edges inserted into the directed SBM Low Cohesion graph is shown.

The experiments with the first intervention, calculated with the FJ-Model, are shown in Figure 4.3a. The only model showing a reduction with this intervention is the FJ-Model, which the intervention was optimized for. There is neither a reduction nor an increase visible in the other models.

In Figure 4.3b the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.05$ are shown. Again, the only model showing a reduction with this intervention is the model it was optimized for, while there is no reduction or increase in polarization visible in the other models.

The experiments with the last intervention, calculated with the BC-Model with $\epsilon = 0.2$ are shown in Figure 4.3c. In this case, the BC-Model with $\epsilon = 0.2$ shows a reduction of 4% with this intervention, while the FJ-Model shows no change in polarization at all and the BC-Model with $\epsilon = 0.05$ shows a slight increase in polarization.



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 20$

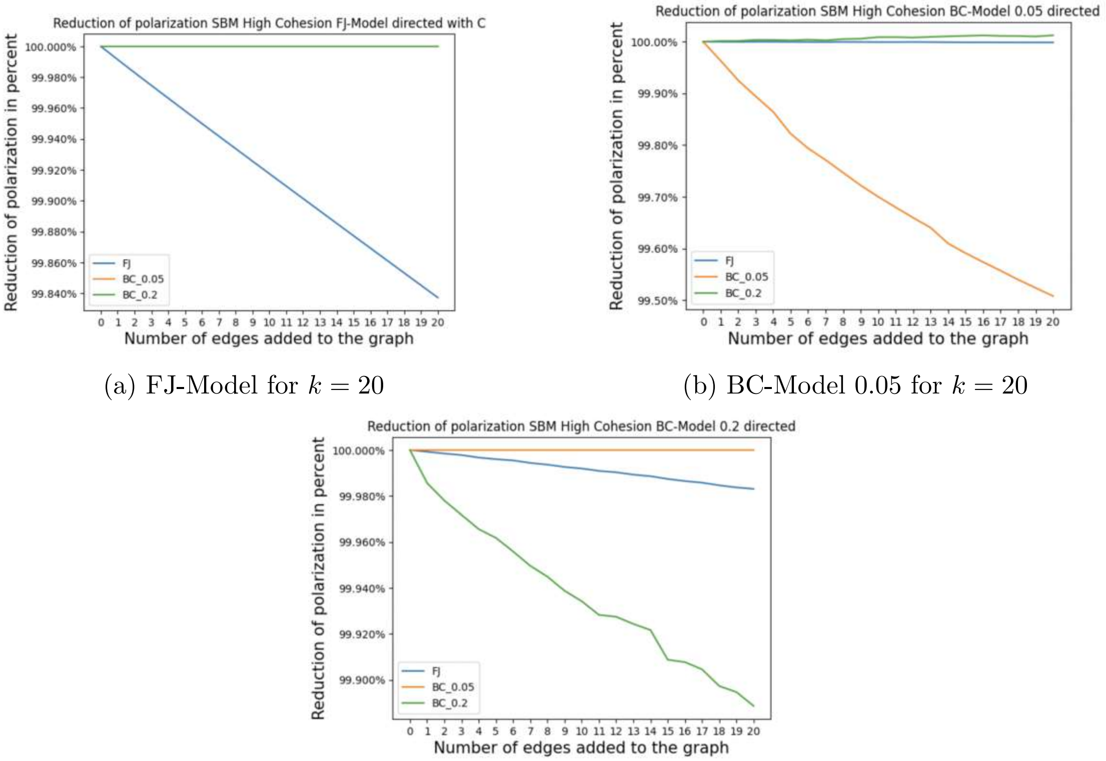
(c) BC-Model 0.2 for $k = 20$

Figure 4.4: Interventions in SBM Low Cohesion undirected calculated with different models

In Figure 4.4, the reduction of polarization in percent depending on the number of edges inserted into the undirected SBM Low Cohesion graph is shown.

As can be seen in Figures 4.4a, 4.4b and 4.4c, the results in the undirected graph are very similar to the results of the directed graph analyzed and described above.

4.3.3 SBM High Cohesion



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 20$

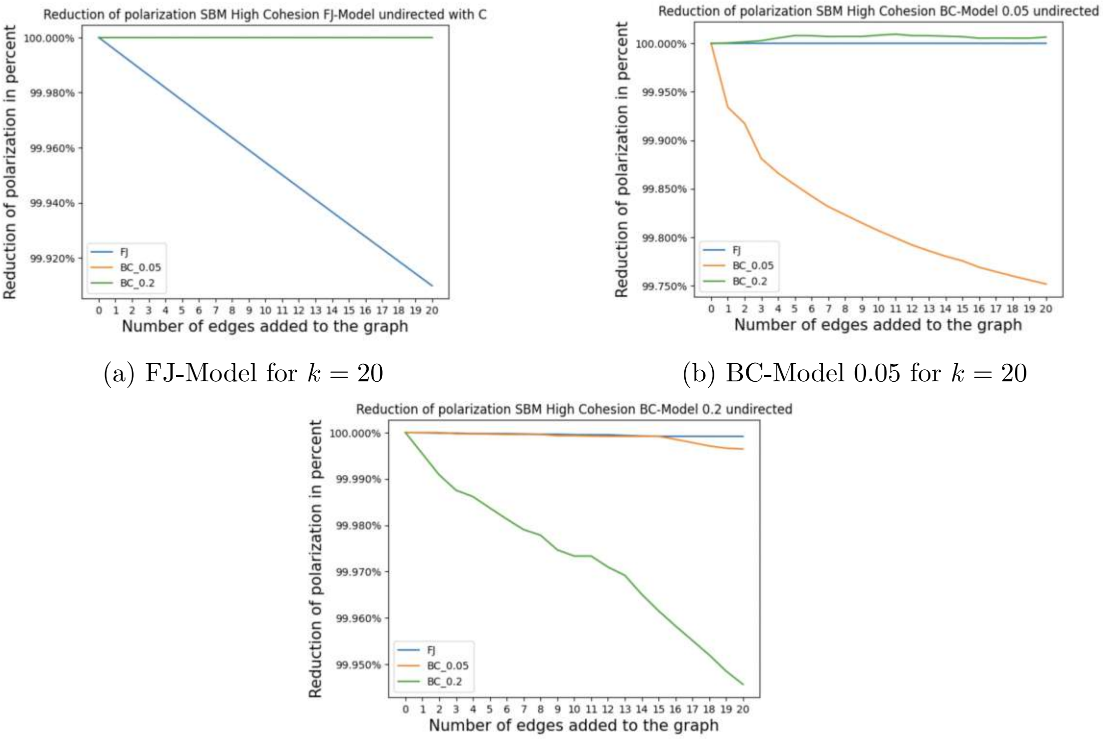
(c) BC-Model 0.2 for $k = 20$

Figure 4.5: Interventions in SBM High Cohesion directed calculated with different models

In Figure 4.5 the reduction of polarization in percent depending on the number of edges added to the directed SBM High Cohesion graph for different interventions is shown. The experiments with the intervention calculated with the FJ-Model are shown in Figure 4.5a. Similar to the SBM Low Cohesion graph, only the FJ-Model achieves a reduction in polarization with this intervention, which it was optimized for, while the other models show no change in polarization at all.

In Figure 4.5b, the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.05$ are displayed. Again, the only model achieving a reduction in polarization is the BC-Model with $\epsilon = 0.05$, which this intervention was optimized for, while the other models show no reduction or increase at all with this intervention.

Lastly, Figure 4.5c shows the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.2$. While the BC-Model with $\epsilon = 0.2$ shows the greatest reduction in polarization, the FJ-Model also shows a small reduction and the BC-Model with $\epsilon = 0.05$ shows no change in polarization at all.



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 20$

(c) BC-Model 0.2 for $k = 20$

Figure 4.6: Interventions in SBM High Cohesion undirected calculated with different models

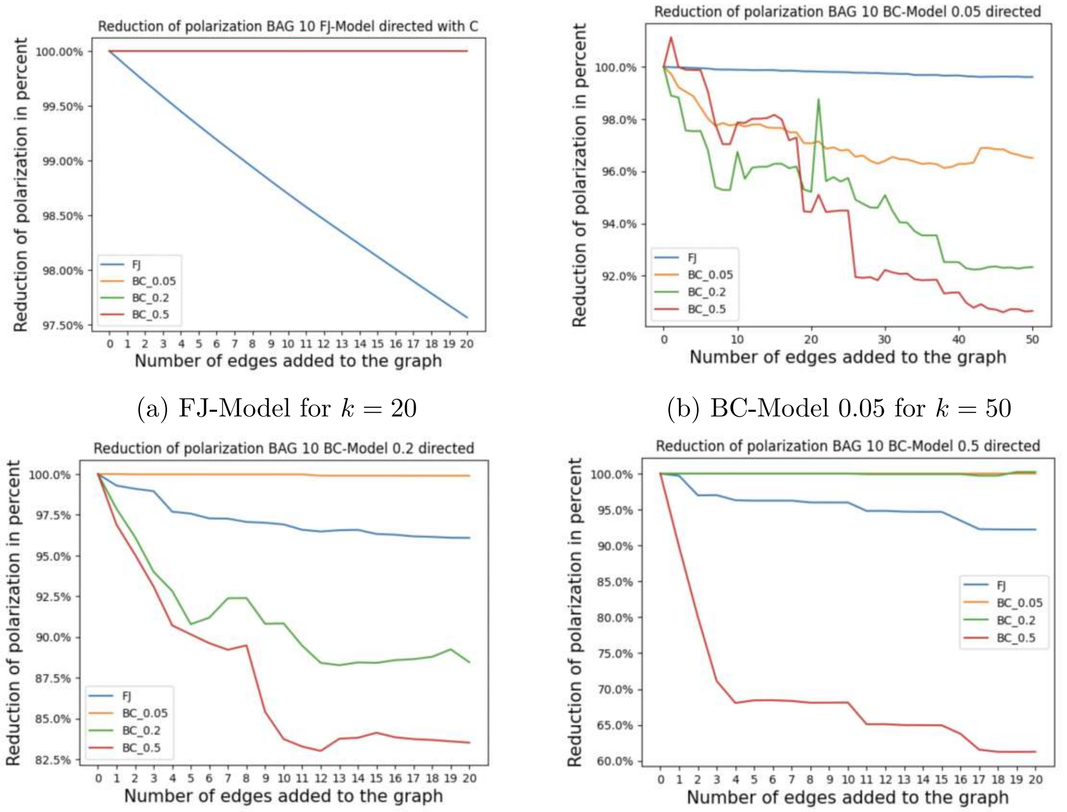
Figure 4.6 shows the reduction of polarization in percent depending on the number of edges inserted into the undirected SBM High Cohesion graph for different interventions. As can be seen in Figures 4.6a, 4.6b and 4.6c the results in the undirected graph are again very similar to the results of the directed graph, as analyzed and discussed above.

4.3.4 Barabasi-Albert-Graph ($m = 10$)

In Figure 4.7, the relative reduction of polarization depending on the number of edges added to the graph with different interventions is shown.

We can see that the intervention calculated with the FJ-Model as seen in Figure 4.7a only reduces polarization in the FJ-Model, whereas there is no reduction or increase visible for any other model.

Moreover, the results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.05$ are shown in Figure 4.7b. Again, the BC-Model with $\epsilon = 0.5$



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 50$

(c) BC-Model 0.2 for $k = 20$

(d) BC-Model 0.5 for $k = 20$

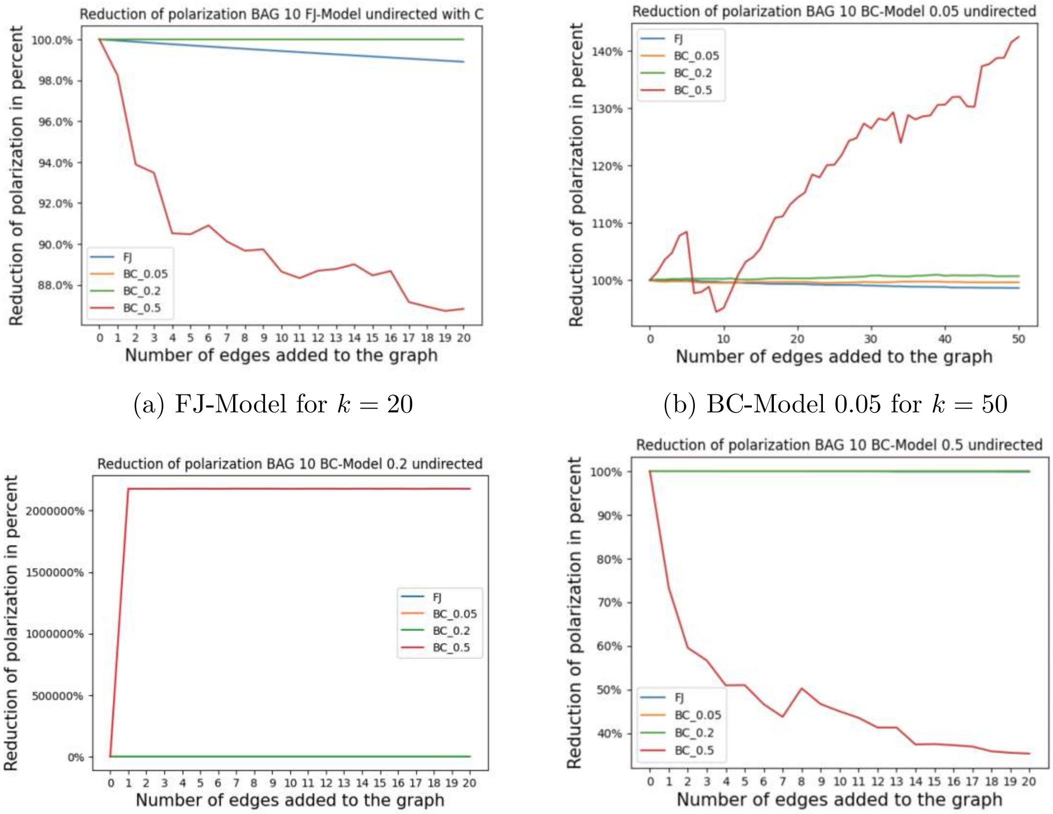
Figure 4.7: Interventions in BAG 10 directed optimized for different models

achieves the largest reduction in polarization, while there is almost no reduction visible in the FJ-Model.

The results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.2$ can be seen in Figure 4.7c. Interestingly, the BC-Model with $\epsilon = 0.5$ achieves the largest reduction of polarization again with this intervention, although the intervention is optimized for the BC-Model with $\epsilon = 0.2$. There is only a small reduction of polarization visible in the FJ-Model and no reduction or increase of polarization in the BC-Model with $\epsilon = 0.05$.

Finally, in Figure 4.7d the results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.5$ are shown. The BC-Model with $\epsilon = 0.5$ achieves the highest reduction of polarization of about 40 % and the FJ-Model achieves a small reduction of polarization as well with this intervention, whereas there is no reduction or increase visible in the other models.

In Figure 4.8, the relative reduction of polarization depending on the number of edges added to the graph using interventions optimized for different models on the undirected



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 50$

(c) BC-Model 0.2 for $k = 20$

(d) BC-Model 0.5 for $k = 20$

Figure 4.8: Interventions in BAG 10 undirected optimized for different models

graphs can be found.

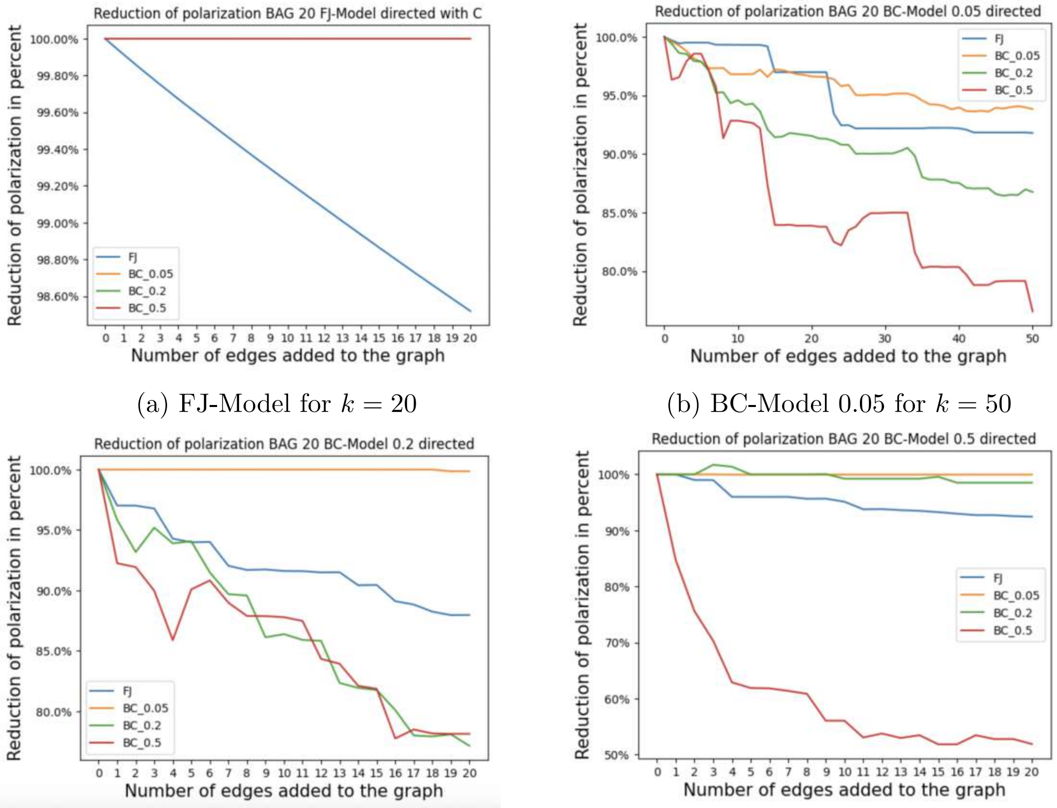
The results of the experiments with the intervention calculated with the FJ-Model can be found in Figure 4.8a. It can be seen that the largest reduction is achieved by the BC-Model with $\epsilon = 0.5$ and there is no reduction or increase in polarization for the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$.

In Figure 4.8b the results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.05$ can be found. While there is a small reduction of polarization in the FJ-Model and the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$, we can also observe an increase in polarization in the BC-Model with $\epsilon = 0.5$.

Furthermore, in Figure 4.8c we can see a very odd behavior of the BC-Model with $\epsilon = 0.5$ - by inserting the first edge into the graph, polarization increases massively and then stays on that high level. This could be due to structural changes in the graph or the initialization of the opinion vector. Moreover, the BC-Model with $\epsilon = 0.5$ already shows very small baseline values of polarization and disagreement, which could also lead to these odd values.

Lastly, in Figure 4.8d, the reduction of polarization depending on the number of edges added to the graph with different interventions is shown. The largest reduction of polarization is achieved by the BC-Model with $\epsilon = 0.5$, while there is no reduction or increase in polarization for all the other models.

4.3.5 Barabasi-Albert-Graph $(m = 20)$



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 50$

(c) BC-Model 0.2 for $k = 20$

(d) BC-Model 0.5 for $k = 20$

Figure 4.9: Interventions in BAG 20 directed optimized for different models

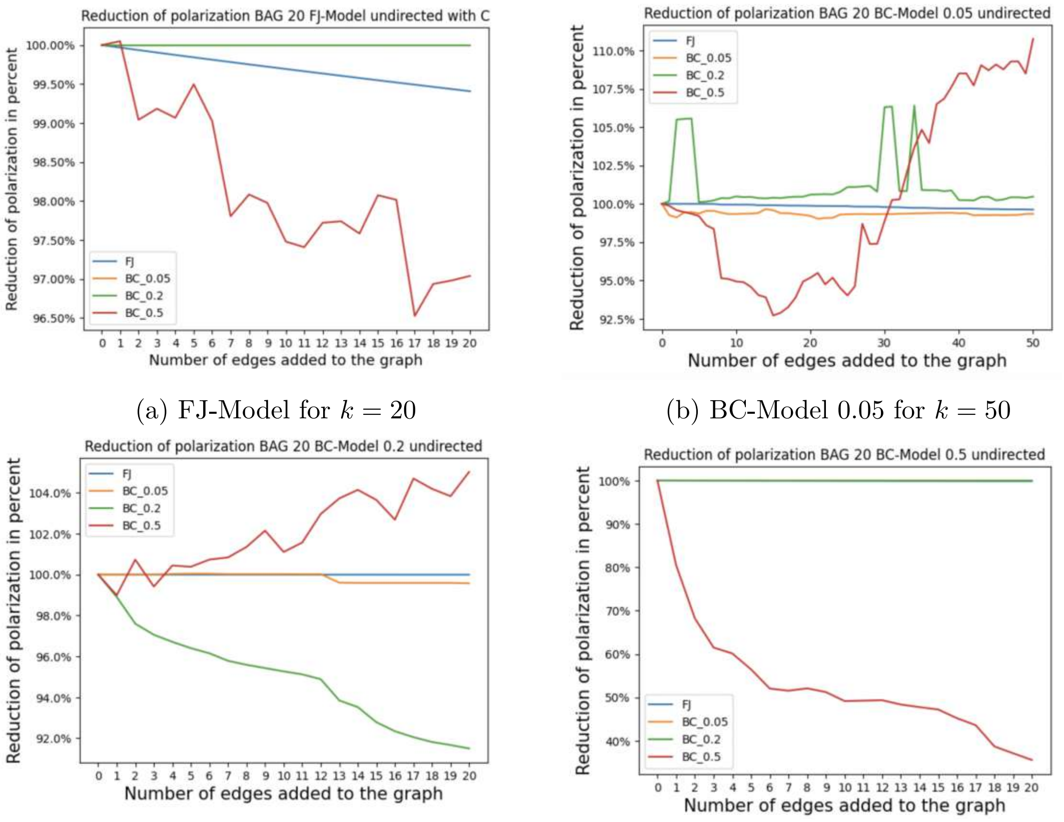
In Figure 4.9, the relative reduction of polarization depending on the number of edges added to the directed graph with different interventions can be found.

In Figure 4.9a we can see that the intervention calculated with the FJ-Model only reduces polarization in the FJ-Model, whereas there is no visible change in polarization in the other models. Furthermore, in Figure 4.9b the largest reduction of polarization is achieved by the BC-Model with $\epsilon = 0.5$, while the smallest reduction of polarization is achieved by the BC-Model with $\epsilon = 0.05$, which the intervention was optimized for.

Moreover, in Figure 4.9c, the results of the experiments with the intervention optimized for the BC-model with $\epsilon = 0.2$ are shown. We can see that the largest reduction of polarization is achieved by the BC-Models with $\epsilon = 0.2$ and $\epsilon = 0.5$. While there is a smaller reduction visible in the FJ-Model, there is no reduction nor increase in polarization visible in the BC-Model with $\epsilon = 0.2$.

Lastly, in Figure 4.9d, the results of the experiments with the intervention optimized for the BC-Model with $\epsilon = 0.5$ are displayed. This model achieves the largest reduction of polarization with this intervention, which is expected, since the intervention was optimized for this model. Additionally, there is a smaller reduction of polarization visible in the FJ-Modell, whereas there is almost no change in polarization visible in the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$.

In Figure 4.10, the relative reduction of polarization depending on the number of edges added to the undirected graph with different interventions is displayed.



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 50$

(c) BC-Model 0.2 for $k = 20$

(d) BC-Model 0.5 for $k = 20$

Figure 4.10: Interventions in BAG 20 undirected optimized for different models

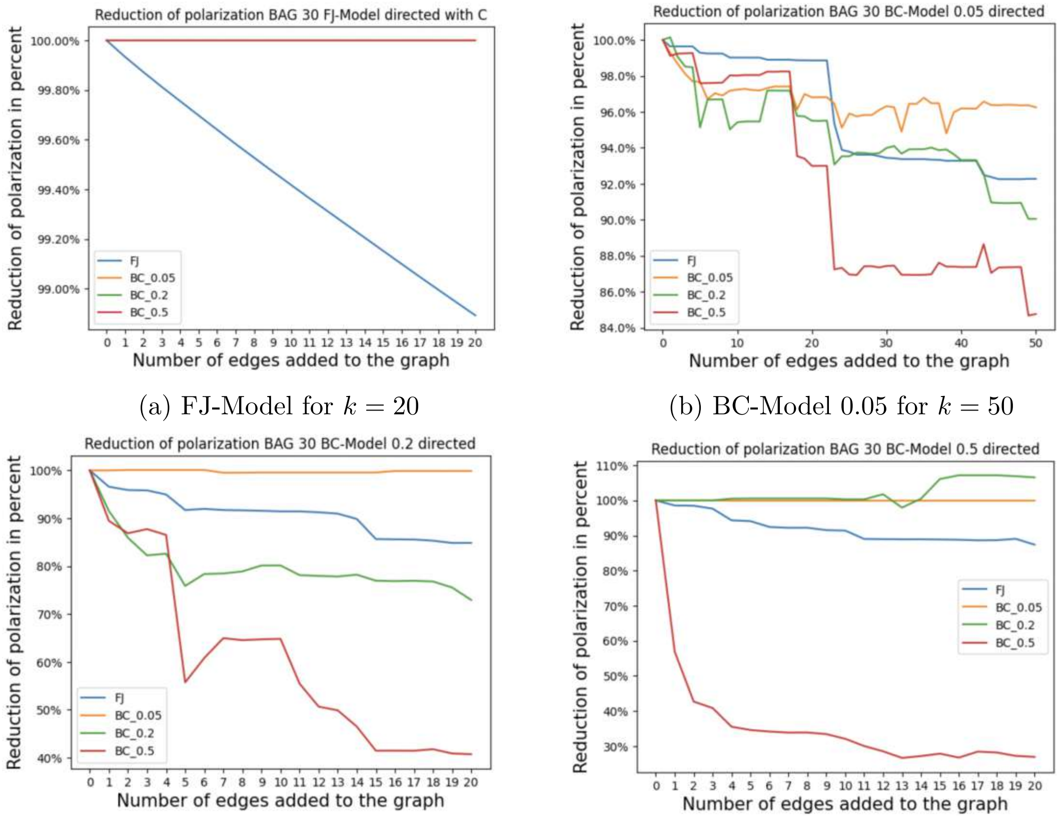
The results of the experiments with the intervention calculated with the FJ-Model on the undirected BAG 20 graph can be found in Figure 4.10a. The largest reduction of polarization with this intervention is achieved by the BC-Model with $\epsilon = 0.5$ and there is a smaller reduction in the FJ-Model. However, these results have to be considered with caution, because the BC-Model with $\epsilon = 0.5$ already has very low baseline values on this dataset.

In Figure 4.10b, the results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.05$ are shown. While the largest reduction of polarization overall is achieved by the model this intervention was optimized for, in the BC-Model with $\epsilon = 0.2$ we can see odd model behavior with this intervention, as the model shows some spikes in polarization and an overall increase. Furthermore, also the BC-Model with $\epsilon = 0.5$ shows first a decline in polarization and later polarization increases again. These results have to be considered with caution because of the low baseline values in this model.

Moreover, in Figure 4.10c the results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.2$ are shown. As can be seen, the largest reduction is achieved with the BC-Model with $\epsilon = 0.2$, which the intervention is optimized for. There is very little to no change in polarization in the FJ-Model and the BC-Model with $\epsilon = 0.05$ and in the BC-Model with $\epsilon = 0.5$, polarization increases.

Lastly, in Figure 4.10d the results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.5$ are shown. With this intervention, the largest reduction in polarization is achieved by the BC-Model with $\epsilon = 0.5$, while there is no reduction or increase shown for all other models.

4.3.6 Barabasi-Albert-Graph $(m = 30)$



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 50$

(c) BC-Model 0.2 for $k = 20$

(d) BC-Model 0.5 for $k = 20$

Figure 4.11: Interventions in BAG 30 directed optimized for different models

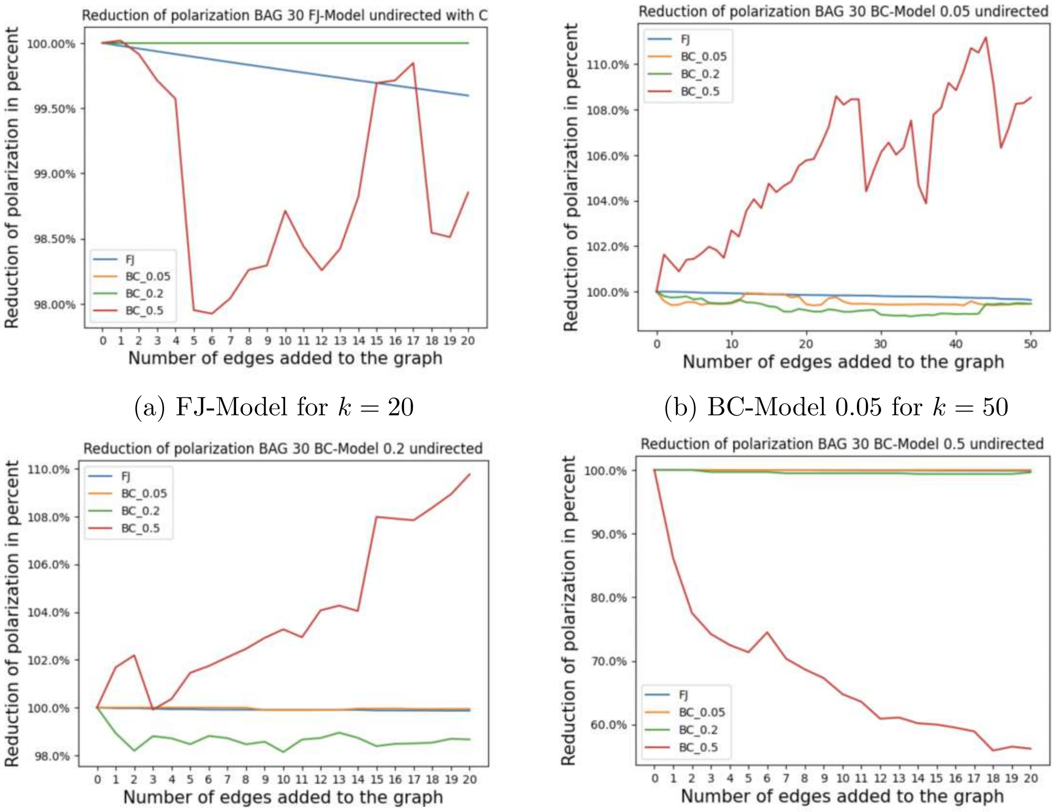
In Figure 4.11 we can see the relative reduction of polarization depending on the number of edges added to the directed BAG 30 graph with different interventions.

The results of the experiments with the intervention calculated with the FJ-Model can be found in Figure 4.11a. As already seen in the other Barabasi-Albert graphs, this intervention only reduces polarization in the FJ-Model, while there is no change in polarization visible in the other models.

In Figure 4.11b the results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.05$ can be seen. This intervention effectively reduces polarization across all models, with the largest reduction reported in the BC-Model with $\epsilon = 0.5$ and the smallest reduction reported in the BC-Model with $\epsilon = 0.05$.

Furthermore, in Figure 4.11c we can see the results of the experiments with the intervention optimized for the BC-Model with $\epsilon = 0.05$. Again, the largest reduction with this intervention is achieved with the BC-Model with $\epsilon = 0.5$, even though the intervention is optimized for a different model, while there are smaller reductions of polarization in the FJ-Model and the BC-Model with $\epsilon = 0.2$. In the BC-Model with $\epsilon = 0.05$, there is neither a reduction nor an increase in polarization visible.

Moreover, in Figure 4.11d the results of the experiments with the intervention calculated with the BC-Model with $\epsilon = 0.5$ can be found. Again, the largest reduction of polarization is achieved by the BC-Model with $\epsilon = 0.5$. There is also a smaller reduction in the FJ-Model visible, but no change in polarization in the BC-Model with $\epsilon = 0.05$. Lastly, with this intervention, there is a small increase in polarization visible in the BC-Model with $\epsilon = 0.2$.



(a) FJ-Model for $k = 20$

(b) BC-Model 0.05 for $k = 50$

(c) BC-Model 0.2 for $k = 20$

(d) BC-Model 0.5 for $k = 20$

Figure 4.12: Interventions in BAG 30 undirected optimized for different models

In Figure 4.12 we can see the relative reduction of polarization depending on the number of edges added to the undirected BAG 30 graph with different interventions.

As can be seen in Figure 4.12a the intervention calculated with the FJ-Model effectively reduces polarization in the FJ-Model, while there is no reduction reported in the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.3$ - similar to the directed case. The largest reduction is reported in the BC-Model with $\epsilon = 0.5$, however, there are some spikes in polarization, which could be due to the already very low baseline values in this model.

In Figure 4.12b, we can see that the largest reduction in polarization with the intervention calculated in the BC-Model with $\epsilon = 0.05$ is achieved by the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$. This intervention seems to trigger a large increase in polarization in the BC-Model with $\epsilon = 0.5$, which again could be due to the very low baseline values, especially in the undirected case of this dataset.

Furthermore, the intervention calculated with the BC-Model with $\epsilon = 0.2$ effectively reduces polarization only in the model it was optimized for, as seen in Figure 4.12c. There is no change in polarization visible in the FJ-Model and the BC-Model with $\epsilon = 0.05$, whereas there is a large increase in polarization in the BC-Model with $\epsilon = 0.5$, which again could be due to the reason given above.

Lastly, the intervention calculated with the BC-Model with $\epsilon = 0.5$ effectively reduces polarization in the same model by about 40%, whereas there is neither a reduction nor an increase in polarization visible for the other models.

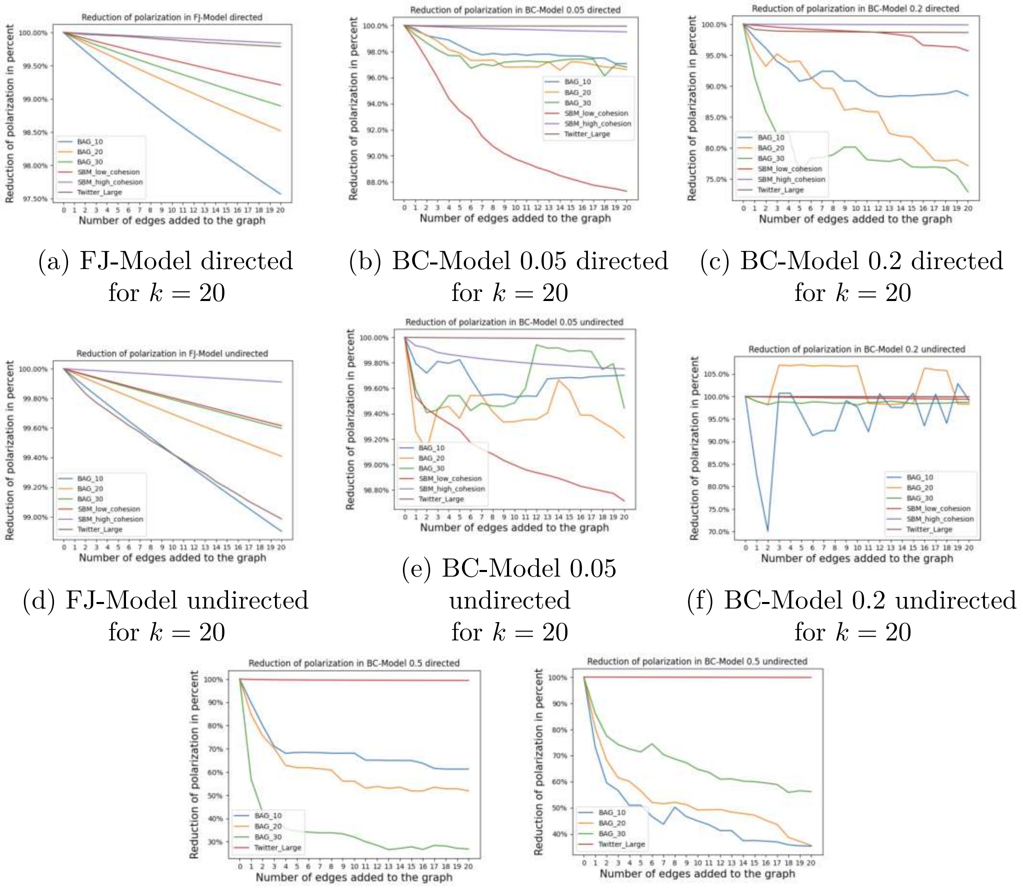
4.4 Comparison of datasets

In this section, the results of the comparison of datasets can be found, which will be discussed further in the Discussion section below.

In Figure 4.13 the reduction of polarization depending on the number of edges added to the graph for different datasets is shown.

The interventions for Twitter Large were obtained with the greedy algorithm specified above on the reduced search space with $k = 100$ in the FJ-Model and $k = 20$ in all other models. However, for reasons of comparability, only the first 20 edges of each intervention were selected. Moreover, the interventions for BAG 10, BAG 20 and BAG 30 were obtained with the greedy algorithm on the full search space defined above with $k = 50$ for the intervention calculated with the BC-Model with $\epsilon = 0.05$ and $k = 20$ for all other models. Again, for reasons of comparability, only the first 20 edges of each intervention were selected. Additionally, the interventions for SBM Low Cohesion and SBM High Cohesion were obtained with the algorithm specified above on the full search space with $k = 20$.

More specifically, in Figure 4.13a, the reduction of polarization with interventions calculated with the FJ-Model on all directed datasets is shown. We can see that the largest reduction in this model is achieved on the graph BAG 10, while the smallest reduction is achieved on SBM high cohesion. Additionally, on the basis of the Barabasi-Albert graphs we can see that the reduction in polarization is smaller, the denser the underlying graph is. Out of the three Barabasi-Albert graphs, the largest reduction is achieved on BAG 10, while the reduction of BAG 20 is in the middle and the smallest reduction is achieved on BAG 30.



(a) FJ-Model directed for $k = 20$

(b) BC-Model 0.05 directed for $k = 20$

(c) BC-Model 0.2 directed for $k = 20$

(d) FJ-Model undirected for $k = 20$

(e) BC-Model 0.05 undirected for $k = 20$

(f) BC-Model 0.2 undirected for $k = 20$

(g) BC-Model 0.5 directed for $k = 20$

(h) BC-Model 0.5 undirected for $k = 20$

Figure 4.13: Comparison of datasets in different models

This behavior can also be observed for the Stochastic Block Models: The reduction of polarization is larger in SBM Low Cohesion than in SBM High Cohesion.

We can observe similar model behavior on the undirected graphs: In Figure 4.13d it can be seen that out of the three Barabasi-Albert graphs, the largest reduction of polarization is achieved by the most sparse graph, BAG 10, whereas the reduction in BAG 20 is in the middle and the smallest reduction is achieved on BAG 30, which is the most dense graph out of the three.

In this case, the second-largest reduction is achieved on the Twitter Large graph, and the smallest reduction in polarization is achieved on SBM High Cohesion.

However, we can see in Figure 4.13b that in the BC-Model with $\epsilon = 0.05$, the order is reversed: Here, out of the three Barabasi-Albert graphs, the largest reduction of polarization is achieved by BAG 30, the second-largest reduction is achieved by BAG

20 and the smallest reduction out of the three is achieved by BAG 10. In this model, the reduction is larger, the more dense the graph is. This can be explained by the fact that in the bounded-confidence model, it is essential that there are enough nodes in the neighborhood with opinions below the ϵ-threshold. In a dense graph, there are more edges, which increases the probability of enough nodes in the neighborhood with an opinion below the threshold, so that polarization can effectively be reduced.

In the BC-Model with $\epsilon = 0.05$, the largest reduction of polarization overall is achieved on the SBM low cohesion graph and the smallest reduction is achieved on the Twitter Large graph. If we run this model on the undirected graphs as shown in Figure 4.13e, we can achieve a similar result as in the directed case.

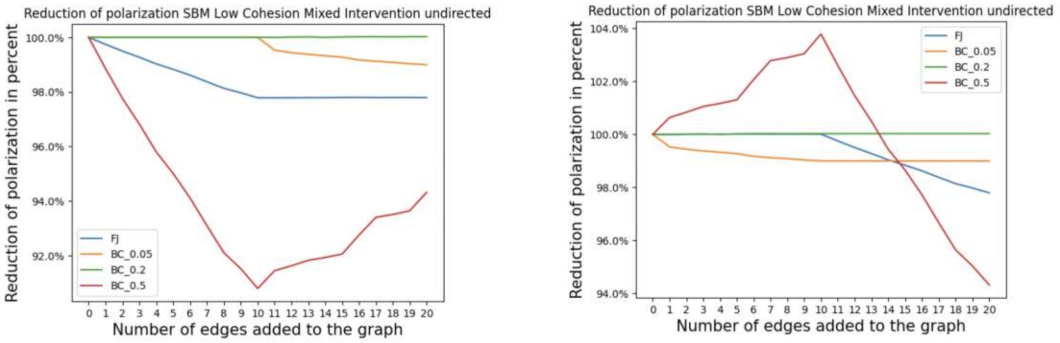
In the BC-Model with $\epsilon = 0.2$ as seen in Figure 4.13c, we can see that the largest reduction of polarization is achieved on BAG 30, whereas the smallest reduction of polarization is achieved on SBM High Cohesion. Comparing the three Barabasi-Albert graphs, again the largest reduction of polarization is achieved on the most dense graph, BAG 30, while the smallest reduction is achieved on the most sparse graph out of the three. This is due to the higher probability of finding enough nodes in the neighborhood with opinions below the ϵ-threshold, in order to effectively reduce polarization.

In the undirected case as seen in Figure 4.13f, the largest reduction of polarization is achieved on BAG 20, while the smallest one is achieved on Twitter Large.

In Figure 4.13g we can see the reduction of polarization in the BC-Model with $\epsilon = 0.5$ on the directed datasets. This model has not been run on SBM Low Cohesion and SBM High Cohesion, since polarization and disagreement are already close to 0 without any intervention on these datasets.

The largest reduction in polarization is achieved on BAG 30 and the smallest reduction is achieved on Twitter Large. The reason for the untypically large reductions in all Barabasi-Albert graphs, but especially in BAG 30, are the already very small values of polarization and disagreement in this model on these datasets. A small absolute change in polarization may lead to a large relative change, since the baseline values are already close to 0.

On the undirected datasets, the largest reduction of polarization is achieved with BAG 10, while the smallest reduction is achieved on Twitter Large, as can be seen in Figure 4.13h. Comparing the three Barabasi-Albert graphs, in this case, the largest reduction is achieved on the most sparse graph out of the three and the smallest reduction is achieved on the most dense graph out of the three. This behavior resembles the behavior of the FJ-Model, which also - out of the three Barabasi-Albert graphs - achieved the largest reduction on BAG 10 and the smallest reduction on BAG 30. This could be due to the fact that the behavior of the model with an ϵ-value of 0.5 assimilates the behavior of the FJ-Model.



(a) Intervention of FJ-Model + Intervention of BC-Model

(b) Intervention of BC-Model + Intervention of FJ-Model

Figure 4.14: SBM Low Cohesion Mixed Intervention

4.5 Mixed Interventions

In this section, the results of the experiments with mixed interventions will be presented, and the discussion of the results can be found in the Discussion section below.

As an illustrative example, two interventions calculated on the undirected SBM Low cohesion graph were concatenated to directly see the impact of inserting edges calculated with one model or the other into the graph on polarization. To also investigate the impact of the order of edges in the intervention, the concatenation was performed in two ways: On the one hand, 10 edges of the intervention calculated with the FJ-Model were concatenated with 10 edges of the intervention calculated with the BC-Model, while on the other hand, the order of the interventions was reversed.

In Figure 4.14 the reduction of polarization with the combined intervention of the intervention calculated with the BC-Model with $\epsilon = 0.05$ and the intervention calculated with the FJ-Model in the undirected SBM Low Cohesion graph can be found. On the left hand side, in Figure 4.14a, we can see the case where the first 10 edges of the intervention stem from the intervention calculated with the FJ-Model and the last 10 edges stem from the intervention calculated with the BC-Model. The edges are also inserted into the graph in this order.

While we can clearly see the reduction of polarization in the FJ-Model with the first half of the intervention, there is no reduction or increase of polarization in the FJ-Model with the second half of the intervention. Additionally, the BC-Model with $\epsilon = 0.5$ responds well to the intervention calculated with the FJ-Model with a reduction in polarization, but inserting the edges calculated with the BC-Model in the graph partly cancels out the reduction and leads to an increase in polarization. For the BC-Model with $\epsilon = 0.05$, we can see no increase or reduction with the edges calculated with the FJ-Model, while the edges optimized to reduce polarization in the BC-Model with $\epsilon = 0.05$ show a distinct effect. In the BC-Model, there is no reduction or increase in polarization visible with either intervention.

Furthermore, on the right hand side in Figure 4.14b we can see the case where the order of the interventions is flipped: Here, the first 10 edges of the intervention stem from the intervention calculated with the BC-Model and the last 10 edges stem from the intervention calculated with the FJ-Model.

While the BC-Model with $\epsilon = 0.5$ also responds with an increase in polarization to the intervention calculated with the BC-Model, polarization decreases in this model with the intervention calculated with the FJ-Model. Moreover, we observe the same model behaviors as with the other order of interventions before: For both the FJ-Model and the BC-Model with $\epsilon = 0.05$, only the edges calculated with the respective models can effectively reduce polarization, whereas there is no increase or reduction in polarization when the edges calculated with a different model are inserted. The BC-Model with $\epsilon = 0.2$ does not respond to either intervention, meaning there is no reduction or increase in polarization at all.

	FJ	BC 0.05	BC 0.2	BC 0.5
Twitter Large	0.998548	0.999588	0.987074	0.997686
SBM Low Cohesion	0.968090	0.914951	0.991851	
SBM High Cohesion	1.006471	0.997713	0.999585	
BAG 10	0.808286	0.984218	0.887443	0.603324
BAG 20	0.761100	0.976843	0.908698	0.571342
BAG 30	0.743648	0.976063	0.782338	0.300964

Table 4.37: Reduction of polarization with mixed intervention on directed graphs

For rest of the experiments with mixed interventions, $\frac{k}{\#models}$ edges of each intervention calculated for a dataset on the full search space were combined to create a new intervention.

The mixed interventions for the datasets BAG 10, BAG 20 and BAG 30 were created as follows: The edges computed with the different models on the full search space using $k = 20$ were combined in alternating order. Consider the intervention calculated with the FJ-Model as A, the intervention calculated with the BC-Model with $\epsilon = 0.05$ as B, the intervention calculated with the BC-Model with $\epsilon = 0.2$ as C and the intervention calculated with the BC-Model with $\epsilon = 0.5$ as D. The mixed intervention was then created as $a_1, b_1, c_1, d_1, a_2, b_2, \ldots,$ where a_1 is the first edge from intervention A, b_1 is the first edge from intervention B and so forth.

The mixed interventions for SBM Low Cohesion and SBM High Cohesion were created as follows: The edges computed with the different models on the full search space using $k = 20$ were combined in alternating order. Consider the intervention calculated with the FJ-Model as A, the intervention calculated with the BC-Model with $\epsilon = 0.05$ as B, the intervention calculated with the BC-Model with $\epsilon = 0.2$ as C. The mixed intervention

	FJ	BC 0.05	BC 0.2	BC 0.5
Twitter Large	0.999862	0.999824	0.999419	0.996268
SBM Low Cohesion	0.994459	0.934452	0.963788	
SBM High Cohesion	0.998860	0.996854	0.999245	
BAG 10	0.984547	0.980564	0.996779	1.015640
BAG 20	0.990065	0.960563	0.848889	0.908550
BAG 30	0.992671	0.986113	0.932186	0.893815

Table 4.38: Comparison of reduction of polarization with model intervention and mixed intervention on directed graphs

was then created as $a_1, b_1, c_1, a_2, b_2, \ldots$, where a_1 is the first edge from intervention A, b_1 is the first edge from intervention B and so forth.

In Table 4.37, the relative reduction of polarization after inserting the edges of the mixed intervention into the directed graph can be seen. While there is a small reduction in polarization across all models for Twitter Large and SBM Low Cohesion, polarization slightly increases in the FJ-Model on SBM High Cohesion. However, in the Barabasi-Albert graphs, there is a significant reduction in polarization, especially in the FJ-Model and the BC-Models with $\epsilon = 0.2$ and $\epsilon = 0.5$.

In Table 4.38, we can see the ratio between the polarization with the intervention calculated with a specific model and the polarization with the mixed intervention. Consider p_{model} the polarization after inserting the edges of the intervention calculated with the respective model and p_{mixed} the polarization after inserting the edges of the mixed intervention, created from all model interventions together in alternating order as described above. The ratio is then defined as $\frac{p_{model}}{p_{mixed}}$. Through this analysis, we can determine which intervention achieves the larger reduction.

We can see that in the FJ-Model, the reduction with the model intervention and the mixed intervention is almost equally large across all datasets. In the BC-Model with $\epsilon = 0.05$, we can see that the model intervention yields a larger reduction than the mixed intervention, but the difference is only marginal. In the BC-Model with $\epsilon = 0.2$ we can see that the reduction of polarization is generally larger with the model intervention, especially in BAG 20. Finally, in the BC-Model with $\epsilon = 0.5$, for BAG 10 the mixed intervention achieves a larger reduction than the model intervention.

In Table 4.39, the relative reduction of polarization with the mixed intervention on the undirected graphs is shown. Similar to the directed case, there is a small reduction of polarization across all models on Twitter Large as well as SBM Low Cohesion and SBM High Cohesion. The mixed interventions achieve a significantly smaller reduction of polarization on the Barabasi-Albert graphs in the undirected case compared to the directed case.

Additionally, there is an unusually high number for the BC-Model with $\epsilon = 0.5$ on BAG

	FJ	BC 0.05	BC 0.2	BC 0.5
Twitter Large	0.999929	0.999946	0.999459	0.999623
SBM Low Cohesion	0.998587	0.991218	0.997801	
SBM High Cohesion	0.999679	0.998311	0.999883	
BAG 10	0.996957	0.998241	0.971870	21761.932243
BAG 20	0.998358	0.993638	1.071488	0.570326
BAG 30	0.998895	0.995406	0.987852	0.697552

Table 4.39: Reduction of polarization with mixed intervention on undirected graphs

10, which could be explained by the already very low baseline values for polarization and disagreement in this case, as mentioned in an example above. This has been investigated further, which revealed that there is one specific edge in the intervention that brings polarization up to this extremely high value, which comes from the BC-Model with $\epsilon = 0.05$.

	FJ	BC 0.05	BC 0.2	BC 0.5
Twitter Large	0.999868	0.999927	0.999354	0.999323
SBM Low Cohesion	0.997294	0.994378	0.995642	
SBM High Cohesion	0.999365	0.998977	0.999562	
BAG 10	0.992039	0.998087	1.019606	0.000016
BAG 20	0.995712	0.999812	0.917147	0.623933
BAG 30	0.997067	0.999258	0.998773	0.805592

Table 4.40: Comparison of reduction of polarization with model intervention and mixed intervention on undirected graphs

In Table 4.40, we can see the ratio between polarization with the model intervention and polarization with the mixed intervention as defined above. Again, we can see that in the FJ-Model, both interventions perform almost equally well across all datasets, as well as in the BC-Model with $\epsilon = 0.05$. In the BC-Model with $\epsilon = 0.2$, there are some differences in the performance of the interventions, namely in BAG 10, the mixed intervention yields a slightly larger reduction of polarization and in BAG 20, the model intervention yields an about 8% larger reduction of polarization than the mixed intervention. Lastly, in the BC-Model with $\epsilon = 0.5$, both interventions perform equally well on Twitter Large, but on all three Barabasi-Albert graphs, the model intervention achieves a larger reduction in polarization. The extremely small ratio in this model on BAG 10 can be explained by the example given above about small baseline polarization values.

The relative change of disagreement with mixed interventions on the directed graphs can be observed in Table 4.41. In the FJ-Model, there is a slight increase in disagreement across all datasets with the mixed intervention, whereas in the BC-Models, there seems to be a reduction in disagreement with the mixed intervention. The largest reduction is

	FJ	BC 0.05	BC 0.2	BC 0.5
Twitter Large	1.000024	0.998834	0.968583	0.996813
SBM Low Cohesion	1.004895	0.902670	0.998011	
SBM High Cohesion	1.002826	0.999966	1.003164	
BAG 10	1.002022	0.981120	0.852820	0.881344
BAG 20	1.000811	0.979462	0.900145	0.851078
BAG 30	1.000488	0.977045	0.674640	0.448058

Table 4.41: Change in disagreement with mixed intervention on directed graphs

achieved by the BC-Model with $\epsilon = 0.5$ on BAG 30.

	FJ	BC 0.05	BC 0.2	BC 0.5
Twitter Large	1.000708	0.999673	1.000193	0.995511
SBM Low Cohesion	1.011979	0.953299	0.960907	
SBM High Cohesion	1.005765	0.991661	0.996870	
BAG 10	1.018297	0.962742	1.003596	0.986831
BAG 20	1.009000	0.954320	0.836198	0.881814
BAG 30	1.005742	0.976898	0.930727	0.878689

Table 4.42: Comparison of change in disagreement with model intervention and mixed intervention on directed graphs

In Table 4.42, the ratio between the disagreement with the model intervention and the disagreement with the mixed intervention on multiple models and across multiple directed datasets can be found. In the FJ-Model, we can see that disagreement increases more with the model intervention than with the mixed intervention across all datasets, however the performance of both interventions is almost equally good.

In the BC-Model with $\epsilon = 0.05$, the increase in disagreement is larger with the mixed intervention than with the model intervention across all datasets. Furthermore, in the BC-Model with $\epsilon = 0.2$, disagreement is higher with the mixed intervention than with the model intervention, except on SBM High Cohesion, where both interventions perform almost equally well. Lastly, in the BC-Model with $\epsilon = 0.5$, disagreement is generally higher with the mixed intervention than in the model intervention.

Lastly, in Table 4.43 the relative change of disagreement on the undirected graphs is displayed. Similar to the directed case, there is a slight increase in disagreement in the FJ-Model across all datasets. However, on the contrary to the results on the directed networks, in the undirected graphs a slight increase in disagreement can also be observed in the BC-Models. Again, there is an unusually high number in the BC-Model with $\epsilon = 0.5$, which could be explained by the already very low baseline value for disagreement in this case.

	FJ	BC 0.05	BC 0.2	BC 0.5
Twitter Large	1.000261	1.000083	1.000098	1.000074
SBM Low Cohesion	1.007089	0.997229	1.004474	
SBM High Cohesion	1.003153	1.000479	1.003436	
BAG 10	1.001977	1.002975	0.978268	23539.454119
BAG 20	1.001070	0.998041	1.063440	0.574752
BAG 30	1.000664	0.998586	0.988287	0.704585

Table 4.43: Change in disagreement with mixed intervention on undirected graphs

	FJ	BC 0.05	BC 0.2	BC 0.5
Twitter Large	1.000599	1.000011	0.999953	0.999849
SBM Low Cohesion	1.015038	0.981257	0.992309	
SBM High Cohesion	1.006388	0.995287	0.996836	
BAG 10	1.015293	0.988770	1.009520	0.000015
BAG 20	1.008031	0.996281	0.924988	0.628112
BAG 30	1.005425	0.999200	0.999010	0.812937

Table 4.44: Comparison of change in disagreement with model intervention and mixed intervention on undirected graphs

Finally, in Table 4.44, we can see the ratio between disagreement with the model intervention and disagreement with the mixed intervention, as defined above. In the FJ-Model, we can see that disagreement is higher with the model intervention than with the mixed intervention across all datasets. In the BC-Model with $\epsilon = 0.05$, both interventions perform almost equally well across all datasets, same as with the BC-Model with $\epsilon = 0.2$. In the BC-Model with $\epsilon = 0.5$, disagreement is higher with the mixed intervention than with the model intervention. The extremely low ratio on BAG 10 can be explained by the extremely high value of disagreement with the mixed intervention.

4.6 Comparison of full and reduced search space

	FJ	BC 0.05	BC 0.2	BC 0.5
SBM Low Cohesion	0.997154	0.872866	0.956791	
SBM High Cohesion	0.999438	0.995064	0.998876	
BAG 10	0.993213	0.970678	0.886595	0.637307
BAG 20	0.988019	0.966217	0.767279	0.468788
BAG 30	0.990947	0.968079	0.735601	0.254029

Table 4.45: Full vs. Reduced Search Space evaluated on polarization on directed graphs

In order to compare the two greedy algorithms developed for this thesis, both methods were evaluated on the datasets SBM Low Cohesion, SBM High Cohesion, BAG 10, BAG 20 and BAG 30. In the tables in this section, the relation between the respective measurement (polarization or disagreement) calculated with the respective intervention in the full search space and the reduced search space can be found.

In this section, the results of comparing the two algorithms can be found, which will be discussed further in the Discussion section below.

The ratio of polarization in the following tables is defined as follows: Let p_{full} be the polarization of z^* of a model after inserting an intervention calculated on the full search space into graph G. Let $p_{reduced}$ be the polarization of z^* of a model after inserting an intervention calculated on the reduced search space into graph G. The ratio displayed in the tables below is defined as $\frac{p_{full}}{p_{reduced}}$.

Analogously, let d_{full} be the disagreement of z^* of a model after inserting an intervention calculated on the full search space into graph G. Let $d_{reduced}$ be the disagreement of z^* of a model after inserting an intervention calculated on the reduced search space into graph G. The ratio of disagreement displayed in the tables below is defined as $\frac{d_{full}}{d_{reduced}}$.

In Table 4.45, the relation between polarization with the intervention calculated on the full search space and the intervention calculated on the reduced search space of the directed graphs is displayed. We can see that the intervention calculated on the full search space achieves a larger reduction in polarization across all models and datasets. More specifically, in the FJ-Model the intervention calculated on the reduced search space performs almost as good as the intervention calculated on the full search space, whereas in the BC-Models, especially with an epsilon-parameter of 0.2 or higher, the interventions calculated on the reduced search space perform significantly worse.

	FJ	BC 0.05	BC 0.2	BC 0.5
SBM Low Cohesion	0.998222	0.989123	0.993813	
SBM High Cohesion	0.999568	0.997455	0.999461	
BAG 10	0.990606	0.997168	0.994114	0.439901
BAG 20	0.994942	0.992101	0.984544	0.340320
BAG 30	0.999473	0.995195	0.986516	0.672733

Table 4.46: Full vs. Reduced Search Space evaluated on polarization on undirected graphs

In Table 4.46 the relation between polarization with the intervention calculated on the full search space and the intervention calculated on the reduced search space on the undirected graphs is displayed. Compared to the directed case discussed above, we can see a very similar result: While the performance of the intervention calculated on the reduced search space is only slightly worse in the FJ-Model and the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$ across all datasets, it performs significantly worse in the BC-Model with $\epsilon = 0.5$.

	FJ	BC 0.05	BC 0.2	BC 0.5
SBM Low Cohesion	1.005822	1.011748	0.998160	
SBM High Cohesion	1.003058	0.998330	1.000769	
BAG 10	1.018459	0.998847	1.000777	0.987379
BAG 20	1.009996	0.994559	1.051358	0.837786
BAG 30	1.000556	0.998649	0.995635	2.163735

Table 4.47: Full vs. Reduced Search Space evaluated on disagreement on directed graphs

Furthermore, in Table 4.47 we can see the relation between disagreement with the intervention calculated on the full search space and the intervention calculated on the reduced search space on the directed graphs.

As can be seen in the table, in the FJ-Model, disagreement increases slightly more with the intervention calculated on the full search space than with the intervention calculated on the reduced search space. There is an unusually high increase with the intervention calculated on the full search space on BAG 30 in the BC-Model with $\epsilon = 0.5$, but this could be due to the already small baseline value of disagreement for this model on this dataset.

	FJ	BC 0.05	BC 0.2	BC 0.5
SBM Low Cohesion	1.010115	0.981110	0.998814	
SBM High Cohesion	1.004399	0.996187	1.000681	
BAG 10	1.018427	0.997707	0.991390	0.437269
BAG 20	1.009623	0.991145	0.983985	0.344033
BAG 30	1.000838	0.999147	0.988757	0.683807

Table 4.48: Full vs. Reduced Search Space evaluated on disagreement on undirected graphs

Finally, in Table 4.48 the relation between disagreement with the intervention calculated on the full search space and the intervention calculated on the reduced search space on the directed graphs can be found.

Again, similar to the directed case analyzed above, there is a slightly higher increase in disagreement in the FJ-Model with the intervention calculated on the full search space than in the intervention calculated on the reduced search space. If we look at the BC-Models with $\epsilon = 0.05$ and $\epsilon = 0.2$, we can see that both interventions performed almost equally well across all datasets, but if we look at the BC-Model with $\epsilon = 0.5$, disagreement seems to be a lot smaller with the intervention calculated on the full search space compared to the intervention calculated on the reduced search space.

CHAPTER 5

Discussion and Conclusion

In this chapter, the results presented in the previous chapter will be analyzed and discussed, insights and lessons learned will be summarized and lastly, conclusions will be drawn.

5.1 Discussion

Regarding the first research question, “What is the impact of the choice of opinion formation model initially?”, it can be said that the baseline values of polarization and disagreement can vary a lot between the different opinion formation models. Some datasets were excluded from further analysis at this stage due to very small polarization values (close to 0, when polarization is always in the range between $[0, 1]$ through normalization), namely the Reddit dataset, or due to scaling issues, namely the Flixster dataset.

In the datasets Twitter Large and SBM Low Cohesion, the FJ-Model achieves the smallest value of polarization, whereas in other datasets, the smallest value is achieved by the BC-Model with $\epsilon = 0.5$. Generally speaking, the FJ-Model seems to achieve a smaller value of polarization initially than the BC-Models, however, polarization can be forced up with the stubbornness parameter c in the FJ-Model. Moreover, the smaller we choose the ϵ-value of the BC-Model, the higher the value of polarization gets, which can be explained logically: The smaller the confidence bound ϵ, the smaller the neighborhood of agents whose opinions are accepted to influence the opinion of an agent in an update. This encourages the formation of clusters, which leads to a higher polarization in the network.

Therefore, the answer to the first research question must be that the choice of opinion formation model initially has a big impact on the values of polarization and disagreement. While the FJ-Model tends to produce lower values of polarization and disagreement, the BC-Models tend to produce higher values of polarization and disagreement. More specifically, the smaller the parameter ϵ, the higher polarization and disagreement tend to be. The differentiation between directed and undirected graph did not seem to have a big impact on polarization and disagreement.

Considering the second research question, "Given the intervention of inserting the same k edges into different models, what is the impact of the choice of opinion formation model?", it can be said that across all datasets and models, the intervention calculated with a specific model achieves the largest reduction of polarization in that specific model. For example, in the FJ-Model, the intervention optimized for that model achieves the largest reduction in polarization, whereas the other interventions achieve either a smaller reduction or no change in polarization at all. An important insight from these experiments is that even after inserting an intervention that was optimized for a different model, polarization does not increase, meaning there is no "harm" done by inserting that intervention.

For some models and datasets, the relative reduction of polarization as defined above seems to be extremely high - for example in the undirected BAG 10 graph with the BC-Model with $\epsilon = 0.5$. These values have to considered with caution though, because - as explained in an example above - for this specific graph and this model, the baseline value of polarization is 0, which is very low, since polarization is always in $[0, 1]$ through normalization.

With respect to disagreement, it can be said that in the FJ-Model, disagreement tends to increase when polarization is reduced, while that is not always the case in the BC-Model. For example, in the directed BAG 20 graph, the intervention calculated with the FJ-Model increases disagreement across all models, while the other interventions slightly decrease disagreement across all models.

Thus, the answer to the second research question is that as expected, considering one specific model, the intervention calculated with that specific model achieves the largest reduction of polarization across all interventions, which is true for both the FJ-Model and the BC-Model. Moreover, the experiments showed that there is no increase in polarization or disagreement when inserting an intervention calculated with a different model, which shows that there is no "damage".

Furthermore, the comparison of datasets revealed that in the FJ-Model, the reduction of polarization is larger, the sparser the underlying graph is, whereas in the BC-Model, the reduction of polarization is larger, the denser the underlying graph is. This could be due to the mechanics of the BC-Model: Since there are in total more edges in a dense graph than in a sparse graph, the probability of having enough nodes in the neighborhood with an opinion below the ϵ-threshold increases, so that polarization can effectively be reduced.

Regarding the third and last research question, "When an algorithm tries to reduce polarization and disagreement across multiple models, to what extent does it still reduce polarization and disagreement?", the experiments conducted in this thesis revealed that the mixed intervention constructed as described in the previous section can still effectively reduce polarization across all datasets. The reduction of polarization of polarization with the mixed intervention is generally larger in the FJ-Model than in the BC-Models, and the smaller we choose ϵ, the smaller the reduction of polarization is. Compared to the interventions calculated with the respective model, the mixed intervention performs almost equally well in the reduction of polarization across all models in both directed and undirected graphs.

Hence, the answer to the third research question is that the mixed interventions perform almost equally well in the reduction of polarization and disagreement as the interventions calculated with the respective model.

Lastly, the comparison of the algorithms developed in this thesis to find interventions on the full and the reduced search space revealed that the interventions calculated on the reduced search space performed almost equally well in the reduction of polarization as the interventions calculated on the full search space. More specifically, in the FJ-Model, both algorithms performed almost equally well across all datasets, whereas in the BC-Models, the algorithm on the full search space yielded better results.

5.2 Conclusion

To conclude, it can be said that the goal of this thesis was to investigate the outcome of interventions on different opinion formation models. To this end, experiments were conducted with the FJ-Model and the BC-Model with different configurations of ϵ on different datasets.

The experiments revealed that in the FJ-Model, the values of polarization and disagreement are initially smaller than in the BC-Models, although polarization in the FJ-Model can be forced up by introducing a stubbornness-parameter c. Moreover, considering one specific model, the intervention calculated with that model achieves the largest reduction of polarization across all interventions, which is true in both the FJ-Model and the BC-Model. What is more, inserting an intervention that was calculated with a different model does not increase polarization and therefore does not do “damage”.

Furthermore, we could see that mixed interventions constructed from the model interventions as described above, can also effectively reduce polarization and disagreement across all models.

Finally, the comparison of the two algorithms developed in this thesis to find interventions on the full and the reduced search space revealed that the interventions calculated on the reduced search space performed almost equally well in the reduction of polarization as the interventions calculated on the full search space.

Overview of Generative AI Tools Used

The tool ChatGPT was used to generate code in Python and C++ for the FJ-Model and the BC-Model. However, the code was adapted to the needs of this thesis and not accepted directly.

Furthermore, CHatGPT was used to help with troubleshooting and finding bugs in the code.

List of Figures

List of Tables

List of Algorithms

Bibliography

[AB02] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[Bra04] Ulrik Brandes. A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25, 03 2004.

[BWV+21] Carmela Bernardo, Lingfei Wang, Francesco Vasca, Yiguang Hong, Guodong Shi, and Claudio Altafini. Achieving consensus in multilateral international negotiations: The case study of the 2015 paris agreement on climate change. *Science Advances*, 7(51):eabg8068, 2021.

[CL16] Michael Crosscombe and Jonathan Lawry. A model of multi-agent consensus for vague and uncertain beliefs. *Adaptive Behavior*, 24(4):249–260, 2016. PMID: 27547020.

[CR20] Mayee F. Chen and Miklos Z. Racz. Network disruption: maximizing disagreement and polarization in social networks, 2020.

[DC19] Media Digital, Culture and Sport Committee. Disinformation and 'fake news': Final report. Eighth report of session 2017–19, House of Commons, British Parliament, 2019.

[DeG74] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69:118–121, 1974.

[DH21] Igor Douven and Rainer Hegselmann. Mis- and disinformation in a bounded confidence model. *Artificial intelligence*, 291:103415, 2021.

[DK11] Igor Douven and Christoph Kelp. Truth approximation, social epistemology, and opinion dynamics. *Erkenntnis*, 75:271–283, 2011.

[DL13] Rogier De Langhe. Peer disagreement under multiple epistemic systems. *Synthese*, 190:2547–2556, 2013.

[DNAW00] Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 03(01n04):87–98, 2000.

[Dou10] Igor Douven. Simulating peer disagreements. *Studies in History and Philosophy of Science Part A*, 41(2):148–157, 2010.

[FIRT15] Paolo Frasca, Hideaki Ishii, Chiara Ravazzi, and Roberto Tempo. Distributed randomized algorithms for opinion formation, centrality computation and power systems estimation: A tutorial overview. *European Journal of Control*, 24:2–13, 2015. SI: ECC15.

[FJ90] Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *The Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

[FJ11] Noah E. Friedkin and Eugene C. Johnsen. *Frontmatter*, page i–vi. Structural Analysis in the Social Sciences. Cambridge University Press, 2011.

[FJB16] Noah E. Friedkin, Peng Jia, and Francesco Bullo. A theory of the evolution of social power: Natural trajectories of interpersonal influence systems along issue sequences. *Sociological Science*, 3(20):444–472, 2016.

[FPTP16] N. E. Friedkin, A. V. Proskurnikov, R. Tempo, and S. E. Parsegov. Network science on belief system dynamics under logic constraints. *Science*, (6310):321–326, 2016.

[Fre56] J.R.P. French. A formal theory of social power. *Psychological Review*, 63(3):181–194, 1956.

[FRTI13] Paolo Frasca, Chiara Ravazzi, Roberto Tempo, and Hideaki Ishii. Gossips and prejudices: Ergodic randomized dynamics in social networks, 2013.

[Heg23] Rainer Hegselmann. Bounded confidence revisited: What we overlooked, underestimated, and got wrong. *Journal of Artificial Societies and Social Simulation*, 26(4), 2023.

[HK02] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence - models, analysis, and simulation. *Journal of Artifical Societies and Social Simulation (JASSS)*, 5(3), 2002.

[J.16] Kunegis J. Flixster network dataset. *KONECT, the Koblenz Network Collection*, 2016.

[JAC05] DIRK JACOBMEIER. Multidimensional consensus model on a barabÁsi–albert network. *International Journal of Modern Physics C*, 16(04):633–646, 2005.

[JFB17] Peng Jia, Noah E. Friedkin, and Francesco Bullo. Opinion dynamics and social power evolution over reducible influence networks. *SIAM Journal on Control and Optimization*, 55(2):1280–1301, 2017.

[KR11] Sascha Kurz and Jörg Rambau. On the hegselmann–krause conjecture in opinion dynamics. *Journal of Difference Equations and Applications*, 17(6):859–876, 2011.

[Kra97] Ulrich Krause. Soziale dynamiken mit vielen interakteuren. eine problemskizze. *Modellierung Und Simulation Von Dynamiken Mit Vielen Interagierenden Akteuren*, pages 37–51, 1997.

[Kur15] Sascha Kurz. Optimal control of the freezing time in the hegselmann–krause dynamics. *Journal of Difference Equations and Applications*, 21(8):633–648, 2015.

[Lor08] Jan Lorenz. *Fostering Consensus in Multidimensional Continuous Opinion Dynamics under Bounded Confidence*, pages 321–334. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[MMT17] Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. Minimizing polarization and disagreement in social networks, 2017.

[PLR06] Alessandro Pluchino, Vito Latora, and Andrea Rapisarda. Compromise and synchronization in opinion dynamics. *The European Physical Journal B - Condensed Matter and Complex Systems*, 50:169–176, 2006.

[PPTF17] Sergey E. Parsegov, Anton V. Proskurnikov, Roberto Tempo, and Noah E. Friedkin. Novel multidimensional models of opinion dynamics in social networks. *IEEE Transactions on Automatic Control*, 62(5):2270–2285, May 2017.

[PTCF17] Anton V. Proskurnikov, Roberto Tempo, Ming Cao, and Noah E. Friedkin. Opinion evolution in time-varying social influence networks with prejudiced agents. *IFAC-PapersOnLine*, 50(1):11896–11901, 2017. 20th IFAC World Congress.

[TN22] Sijing Tu and Stefan Neumann. A viral marketing-based model for opinion dynamics in online social networks. In *Proceedings of the ACM Web Conference 2022*, pages 1570–1578, New York, NY, USA, 2022. ACM.

[WXJ24] Lingfei Wang, Yu Xing, and Karl H. Johannsson. On final opinions of the friedkin-johnsen model over random graphs with partially stubborn community. *IEEE 63rd Conference on Decision and Control (CDC)*, 2024.

[ZBZ21] Liwang Zhu, Qi Bao, and Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2072–2084. Curran Associates, Inc., 2021.

[ZNGG24] Tianyi Zhou, Stefan Neumann, Kiran Garimella, and Aristides Gionis. Modeling the impact of timeline algorithms on opinion dynamics using low-rank updates. 2024.

[ZW22] Qinyue Zhou and Zhibin Wu. Multidimensional friedkin-johnsen model with increasing stubbornness in social networks. *Information Sciences*, 600:170–188, 2022.