Impact of Censorship in Social Media

Kacper Krasowski

Abstract

This paper examines the impact of social media censorship on polarization, disagreement, and internal conflict using the Friedkin-Johnsen opinion dynamics model. The study assumes a network with two types of agents characterized by opposing viewpoints. Censorship is modeled by banning extreme opinions, and its effects are analyzed. Results show that higher censorship increases polarization, but decreases internal conflict. The effect on disagreement depends on network size and connectivity; it decreases in smaller networks but initially increases in larger networks before eventually declining. For certain types of networks, the optimal censoring was found. The effects of censorship on the indices, turn out to be greater in the network with low homophily.

Advisor: Pau Milán
Affiliation: QEM-IDEA

1 Introduction

Even though social media platforms have become highly influential in shaping real-world dynamics, legislation and government control are almost nonexistent. This suggests that policymakers will likely become interested in social media regulations as evidenced by the recent release of the Digital Services Act. A particularly interesting aspect is the issue of freedom of speech on social media. This naturally raises questions about the effects of censorship, specifically, *what are the implications of limiting freedom of speech on social media for overall welfare? Is there a level of censorship that maximizes welfare?* Additionally, different topics and discussions online exhibit varying degrees of homophily between opposing groups. For instance, the discussions about climate change, opposing sides frequently engage with each other. In contrast, debates about abortion laws often occur within echo chambers. Therefore, we also aim to explore *how the homophily of a network impacts the effects of censorship*.

We address these questions by using the Friedkin-Johnsen opinion dynamics model. In this model each agent begins with an innate opinion, which evolves through communication with other agents until reaching equilibrium, at which all opinions remain unchanged. With higher intensity of communication, the opinions evolve faster. Our model assumes two types of agents, each representing opposing viewpoints, characterized by distinct initial opinion distributions. The intensity of communication between agents is different within and across these types. Censorship is introduced by removing agents with extreme opinions from the social media platform. The timing of the model is as follows: First, the network administrator learns the type of each agent and the intensities of their communication. Then, the censorship level is determined. Based on this level, certain agents are banned, retaining their innate opinions, while the remaining agents continue to interact until they reach equilibrium. We explore, from the perspective of setting the censorship level, how it affects polarization, disagreement, and internal conflict at equilibrium. Additionally, we construct a welfare index derived from these indices to provide a comprehensive assessment.

Our findings indicate that, from the administrator's perspective, higher censorship results in increased polarization and reduced internal conflict. The impact on disagreement varies depending on the network parameters. In small and poorly connected networks, disagreement decreases with higher censorship. On the other hand, in larger or highly connected networks, disagreement

initially increases with censorship, but eventually starts to decrease after a certain point. Furthermore, for certain structures of a network, we identify an optimal level of censorship that maximizes welfare. Depending on the parameters and the structure of the welfare function, it may be optimal to ban some, all, or none of the opinions. Lastly, in network structures with extremely high homophily, censorship has minimal impact on the indices. In networks with extremely low homophily, the indices are highly affected.

As of December 2022, the number of monthly active users worldwide on Twitter was 368 million, as shown in Dixon (2023). Social media usage is constantly growing, and it is becoming increasingly influential in real-life opinion formation (A. Smith and Anderson (2018)). The following figures, from Ortiz-Ospina (2019), present the growth and total usage of social media platforms:

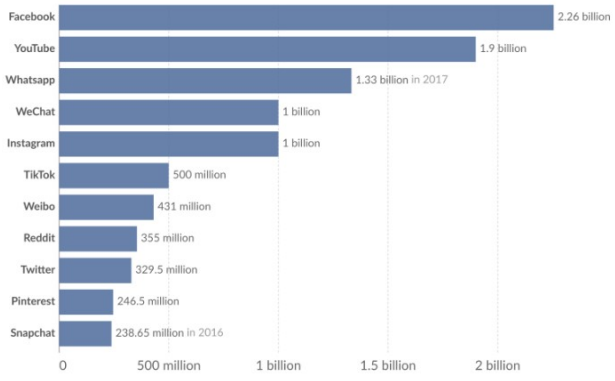

Figure 1: Social Media Usage

Number of people using social media platforms in 2018, source: OurWorldInData.org

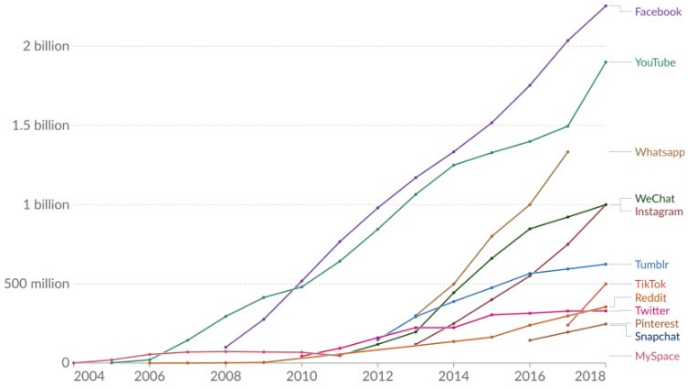

Figure 2: Growth of Social Media Usage

Number of people using social media platforms from 2004 to 2018, source: OurWorldInData.org

To the best of our knowledge, the quantitative investigation of censorship and its effects on society has not been previously proposed or explored, despite its pressing relevance. This gap in the research presents significant opportunities for further study. The combination of the topic's importance and the lack of existing research creates the necessity of this paper. Our contribution to the literature involves proposing a way of modeling the censorship in social media and examining its impact.

1.1 Literature Review

The following reviews the associated literature regarding opinion dynamics, freedom of speech in social media, and the impact of social media on welfare literature.

Opinion Dynamics

As noted in Xia, Wang, and Xuan (2011), the literature regarding opinion dynamics has been growing fast, and there has been an increasing interest in applying the tools for analysis of various disciplines. Among these we can find such disciplines as politics, which were analyzed in e.g. Acemoglu and Ozdaglar (2011), Baer (2016), and Braha and De Aguiar (2017); viral marketing, analyzed in Quattrociocchi, Caldarelli, and Scala (2014); climate change, analyzed in McCright and Dunlap (2011) and healthcare, where the analysis can be found in e.g. Holone (2016), and K. P. Smith and Christakis (2008).

In past decades there have been many new models trying to explain the dynamics of opinions. In Clifford and Sudbury (1973), the authors proposed the Voter Model, which concerns the social discrete choice such as voting on two candidates. DeGroot (1974) proposed a continuous opinion dynamic model, in which the opinions are formed as a weighted average of opinions of neighbours of the agents. Deffuant et al. (2000) created the Bounded Confidence Model, which takes into account the fact that opinion would not be influenced by another agent if the difference between the two agents' opinions is larger than a given threshold. On the other hand, Friedkin and Johnsen (1997) added the stubbornness by assuming that the agents, with some weight, tend to stick to their innate opinions. The last one has been widely used to analyze the opinion dynamics in social media.

The closest literature to this topic lies in opinion dynamics in social media. Among the very wide literature, there are two papers which had a significant impact on this thesis. Cameron Musco, Christopher Musco, and Tsourakakis (2018) introduced a very compact definition of polarization and disagreement in social media. They focus on the analysis of Twitter and Reddit in a utopic setting, in which the owners of social media would minimize the sum of Polarization and Disagreement over opinions and over graphs. On the other hand, Chitra and Christopher Musco (2020) introduce an idea of a Network Administrator, who minimizes disagreement over a graph. They

show that such minimization leads to higher polarization in the network.

Freedom of Speech in Social Media

Segado-Boj and Campo Lozano (2020) focus on three aspects of social media: freedom of information, privacy, and free speech. In the paper, they perform a qualitative analysis of these components. As potential problems regarding freedom of speech, they list: boundaries of free speech and arbitrary censorship, which means that *companies have acquired the potential power to control the information flow as well as hide or silence issues*. Alkiviadou (2019) looks into the Code of Conduct between the European Commission and the IT Companies, and its implications on freedom of speech, and hate speech. Dehghan (2018) performs a network analysis of freedom of speech, and polarization based on a case study of Australia's Racial Discrimination Act (RDA), which restricted freedom of speech in the public sphere.

Impact of Social Media on Welfare

The literature regarding the implications of Social Media usage mostly focuses on in-real-life indicators. Allcott et al. (2020) studies social media's welfare impact based on subjective well-being and analyzes substitutional patterns. Bao, Liang, and Riyanto (2021) look into the effects of social media browsing and social media communication on users' life satisfaction. There also exists a broad literature emphasizing the welfare of children in the context of social media usage e.g. M. Sage and T. Sage (2016); Breyette and Hill (2015).

Roadmap

The paper is organized as follows: Section 2 outlines the methodological contributions and summarizes all the necessary components of the problem. Section 3 presents the results and provides the underlying intuitions. Finally, Section 4 concludes the paper and suggests directions for future research.

2 Methodology

To investigate the impact of censorship on polarization, disagreement, and internal conflict, we develop a simple model describing opinion dynamics. In such models, opinions represent the degree of agreement with a certain statement, such as "there should be a wall built between Mexico and USA" or "abortion should be legal". These opinions are encoded as real values within the continuous interval $[-1,1]$, where an opinion of 1 signifies complete agreement, -1 signifies complete disagreement, and 0 signifies neutrality on the topic.

We assume that there are n agents, divided into two equal groups, each consisting of $\frac{n}{2}$ agents. These groups represent opposite initial opinions $\bar{s}$ on a certain topic, which can be seen as analogous to left-wing and right-wing perspectives. Each group is characterized by distinct distributions of opinions. Left wing is characterized by F_L with support $[-1,0]$, mean $-\mu$ and variance σ^2 and right wing is characterized by F_R with support $[0,1]$, mean μ and variance σ^2. These distributions are symmetric to each other at 0, which makes the opinions to be mean-centered. Additionally, the agents communicate with each other, within the type with a intensity p, and across types with a intensity q. Intensities are set to be in $[0,1]$.

The censorship is introduced by either policy maker or social media platform owner, both referenced as the network administrator throughout the remainder of this paper. The network administrator picks a level $c \in [0,1]$, which is a censor point such that, if an agent has an opinion outside of $[-c,c]$ threshold, it gets banned from the social media platform. Being banned means that an agent will keep his current opinion forever and will neither be influenced by others, nor influence anybody else's opinion. It is important to note that, the smaller the c, the higher the censorship, as the interval for accepted opinions gets smaller.

A priori, Network administrator knows the type of each agent, as he could infer it from previous online activity, distribution F_L and F_R, and intensities p and q.

2.1 Network

We assume that the network is represented by a undirected graph $G(E,V,w)$, where each node is an agent, each edge is an indicator of agents communicating, and weight is the intensity of communication among agents. There is adjacency matrix, which represents the strength of communication

across agents, $\mathbf{A}$[1] and laplacian of the graph $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is a diagonal matrix of degrees[2]. In this model the agents within the same group communicate with each other with intensity p and the agents in the across groups communicate with intensity q. Thus, adjacency matrix and laplacian of the graph are:

$$\mathbf{A} = \begin{bmatrix} p\mathbf{J}_{\frac{n}{2}} & q\mathbf{J}_{\frac{n}{2}} \\ q\mathbf{J}_{\frac{n}{2}} & p\mathbf{J}_{\frac{n}{2}} \end{bmatrix} \tag{2.1}$$

$$\mathbf{D} = n\frac{p+q}{2}\mathbf{I}_n \tag{2.2}$$

$$\mathbf{L} = n\frac{p+q}{2}\mathbf{I}_n - \begin{bmatrix} pJ_{\frac{n}{2}} & qJ_{\frac{n}{2}} \\ qJ_{\frac{n}{2}} & pJ_{\frac{n}{2}} \end{bmatrix} \tag{2.3}$$

Where $\mathbf{J}_{\frac{n}{2}}$ is a matrix full of ones of size $\frac{n}{2}$ by $\frac{n}{2}$ and $\mathbf{I}_n$ is an identity matrix of size n by n.

2.2 Friedkin-Johnsen Model

In order to investigate the effects of censorship, the Friedkin-Johnsen Opinion Dynamics model is employed. In this model, each agent starts with an innate opinion $\bar{s}_i$, which is its first period opinion and the opinion that it tends to stick to with some degree of stubbornness. The opinions $\bar{z}_i^{(t)}$ evolve in a discrete time setting, with $t = 0, 1, 2, ...$ Each period, agents communicate with their neighbours and set new opinion as a weighted average of the neighbours' opinions and its innate opinion, namely:

$$\bar{z}_i^{(t)} = \frac{\bar{s}_i + \sum_j a_{ij}\bar{z}_j^{(t-1)}}{1 + d_i} \tag{2.4}$$

, where a_{ij} is the intensity of edge (i, j), d_i is the degree of the agent i, $\bar{z}_i^{(t)}$ is the opinion of agent i in period t, and $\bar{s}_i$ is the innate opinion of agent i.

[1] Self-loops are allowed for simplicity, but they do not change the results in any way.

[2] Sum of connectives of an agent.

Given that, we can define the dynamics in the matrix form:

$$\bar{\mathbf{z}}^{(\mathbf{t})} = (\mathbf{I} + \mathbf{D})^{-\mathbf{1}}(\bar{\mathbf{s}} + \mathbf{A}\bar{\mathbf{z}}^{(\mathbf{t-1})}) \tag{2.5}$$

,where $\mathbf{A}$ is the adjacency matrix, $\mathbf{D}$ is the diagonal matrix of degrees, $\bar{\mathbf{z}}^{(\mathbf{t})}$ is the vector of opinions in period t, and $\bar{\mathbf{s}}$ is the vector of innate opinions.

2.2.1 Equilibrium

For the purpose of investigating the effects of censorship, the equilibrium opinions are the point of interest. The equilibrium opinions $\bar{\mathbf{z}}$ are such that, they do not change after certain period. Given a graph G and innate opinions $\bar{\mathbf{s}}$, the opinions converge to the equilibrium values:

$$\bar{\mathbf{z}} = \lim_{t \to \infty} \bar{\mathbf{z}}^{(\mathbf{t})}$$

Thus, the equilibrium opinions are given by:

$$\bar{\mathbf{z}} = (\mathbf{I} + \mathbf{L})^{-\mathbf{1}}\bar{\mathbf{s}}$$

2.3 Welfare Measures

To investigate the effect of censorship, we will focus on three indexes, specifically: polarization, disagreement, and internal conflict. It is worth noticing that, the higher each index, the worse it is for society. Therefore, in terms of these indices, welfare should be defined as negative weighted sum of those terms. As mentioned at the beginning of this section, the opinions are mean-centered, indicating how much they differ from the average opinion. This approach will be useful for defining the forthcoming indices.

2.3.1 Polarization

Polarization reflects the extent of differences in opinions in a society, measuring how far apart these opinions are from one another. Higher polarization indicates greater variance in opinions. For the purposes of this paper, polarization is defined as the variance of the set of equilibrium opinions. Given that the opinions are mean-centered, this variance can simply be expressed as the sum of their squares.

Polarization

Variance of a set of opinions.

$$\mathcal{P} = \sum_{i=1}^{n} \bar{z}_i^2 = \bar{\mathbf{z}}^{\mathbf{T}} \bar{\mathbf{z}} = \bar{\mathbf{s}}^{\mathbf{T}} (\mathbf{I} + \mathbf{L})^{-\mathbf{2}} \bar{\mathbf{s}}$$

With increasing censorship we could expect the polarization to increase, as people would communicate less, and in equilibrium their opinions would be more distant.

2.3.2 Disagreement

Disagreement demonstrates how much opinions differ across connected agents in a society, reflecting the overall level of arguments. The higher the disagreement, the more intense the arguments, or more arguments in general. Thus, disagreement can be defined as the sum of the squared differences in equilibrium opinions of connected agents, weighted by the intensity of their connections.

Disagreement

How much opinions differ in a network.

$$\mathcal{D} = \sum_{i=1}^{n} \sum_{j \in N(i)} a_{ij} (\bar{z}_i - \bar{z}_j)^2 = \bar{\mathbf{z}}^{\mathbf{T}} \mathbf{L} \bar{\mathbf{z}}$$
$$= \bar{\mathbf{s}}^{\mathbf{T}} (\mathbf{I} + \mathbf{L})^{-\mathbf{1}} \mathbf{L} (\mathbf{I} + \mathbf{L})^{-\mathbf{1}} \bar{\mathbf{s}}$$

The effect of censorship on disagreement is complex to predict. Banning individuals would reduce the overall number of conversations, potentially decreasing disagreement. However, with fewer interactions, opinions may become less aligned, potentially increasing the distance between them. Thus, censorship could simultaneously decrease and increase disagreement.

2.3.3 Internal Conflict

Internal conflict measures the extent to which each agent's opinion deviates from their innate beliefs. It indicates how far each agent has shifted from their core values. A higher internal conflict signifies greater evolution in opinions. While it may initially seem less important than the previous indices, higher internal conflict can lead to increased uncertainty, instability, and stress among individuals as they struggle to reconcile their evolving views with their original beliefs. It can

be quantified as the sum of the squared differences between each agent's innate opinion and their equilibrium opinion[3].

Internal Conflict

How much equilibrium opinions differ from innate opinions.

$$\mathcal{IC} = \sum_{i=1}^{n} (\bar{z}_i - \bar{s}_i)^2 = \bar{\mathbf{s}}^{\mathbf{T}} (\mathbf{I}+\mathbf{L})^{-\mathbf{1}} \mathbf{L}^{\mathbf{2}} (\mathbf{I}+\mathbf{L})^{-\mathbf{1}} \bar{\mathbf{s}}$$

With increasing censorship we could expect the internal conflict to decrease, as people would communicate less, and the equilibrium opinions would be less distant from the innate opinions.

It is important to note that each of the indices is a quadratic form of $\bar{\mathbf{s}}$, which for each index is defined by different matrix $f_{\mathcal{I}}(\mathbf{L})$, namely $f_{\mathcal{P}}(\mathbf{L}) = (\mathbf{I}+\mathbf{L})^{-2}$,
$f_{\mathcal{D}}(\mathbf{L}) = (\mathbf{I}+\mathbf{L})^{-1}\mathbf{L}(\mathbf{I}+\mathbf{L})^{-1}$, $f_{\mathcal{IC}}(\mathbf{L}) = (\mathbf{I}+\mathbf{L})^{-1}\mathbf{L}^2(\mathbf{I}+\mathbf{L})^{-1}$, and thus each index can be written as:

$$\begin{aligned}
\mathcal{P} &= \bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{P}}(\mathbf{L}) \bar{\mathbf{s}}, \\
\mathcal{D} &= \bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{D}}(\mathbf{L}) \bar{\mathbf{s}}, \\
\mathcal{IC} &= \bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{IC}}(\mathbf{L}) \bar{\mathbf{s}}
\end{aligned}$$

Which will become useful later on.

2.3.4 Welfare

As mentioned earlier, the higher each index, the worse it is for the society. Thus, if we want to define welfare in terms of those indices we should treat them as disutility, thus the index is a negative sum of these. As it is not known how each of the indices is important for the society, the welfare function will be parameterized by $\alpha \in [0,1]$ and $\beta \in [0, 1-\alpha]$, namely:

[3] See appendix for detailed derivation.

Welfare

Weighted negative sum of all polarization, disagreement and internal conflict.

$$\mathcal{W} = -[\alpha\mathcal{P} + \beta\mathcal{D} + (1-\alpha-\beta)\mathcal{IC}]$$

2.4 Censorship

In the scenario of interest, the network administrator must select the censor point c before the evolution of opinions begins. The process unfolds as follows: First, the network administrator learns the distributions F_L and F_R, along with the parameters p and q. Next, the network administrator chooses the censor point c. With this chosen c, agents with opinions outside of $[-c,c]$ threshold get banned, and other agents communicate and reach equilibrium. The point of interest is to understand the perspective of the network administrator. It is important to note that the lower the censor point, the less is allowed to say, thus the higher the censorship is.

Since the network administrator only knows the distributions, she can only use expectations. Given that the distributions F_L and F_R are symmetrical around 0 and that banning is also symmetrical around 0, each agent has an identical probability of being banned. The left type has a probability $1-F_L(-c)$ of being banned, and the right type has a probability $1-F_R(c)$ of being banned. These probabilities are identical and will be denoted as $1-F(c)$ for future reference. Therefore, from the perspective of the network administrator, each agent has a probability of $F(c)$ of remaining on the platform. Thus, with censorship, the equations 2.1, 2.2, and 2.3 become:

$$\mathbf{A} = F(c)\begin{bmatrix} p\mathbf{J}_{\frac{n}{2}} & q\mathbf{J}_{\frac{n}{2}} \\ q\mathbf{J}_{\frac{n}{2}} & p\mathbf{J}_{\frac{n}{2}} \end{bmatrix} \tag{2.6}$$

$$\mathbf{D} = F(c)n\frac{p+q}{2}\mathbf{I}_n \tag{2.7}$$

$$\mathbf{L} = F(c)\left(n\frac{p+q}{2}\mathbf{I}_n - \begin{bmatrix} p\mathbf{J}_{\frac{n}{2}} & q\mathbf{J}_{\frac{n}{2}} \\ q\mathbf{J}_{\frac{n}{2}} & p\mathbf{J}_{\frac{n}{2}} \end{bmatrix}\right) \tag{2.8}$$

3 Results

To illustrate the intuitive relationship between censorship and the indices, let us consider a simple example:

Example 1:

Let us consider $n = 6$, $p = 0.3$, $q = 0.1$. Agents 1, 2, 3 belong to the left group, and 4, 5, 6 belong to right group. This set-up could be represented by the following figure, with their opinions that can be seen inside of the nodes[4]:

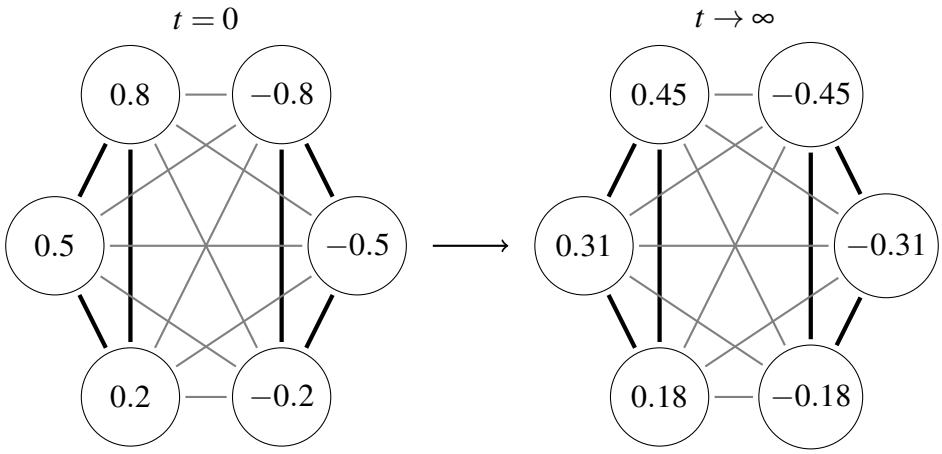

Figure 3: Initial network with all agents for $t = 0$ and $t \to \infty$

In this case, all of the agents evolve with their opinions to equilibrium. Now lets consider the scenario in which the two most extreme agents get banned. This case could be represented by the following figure:

[4]Opinions rounded to the second decimal place in case of equilibrium opinions.

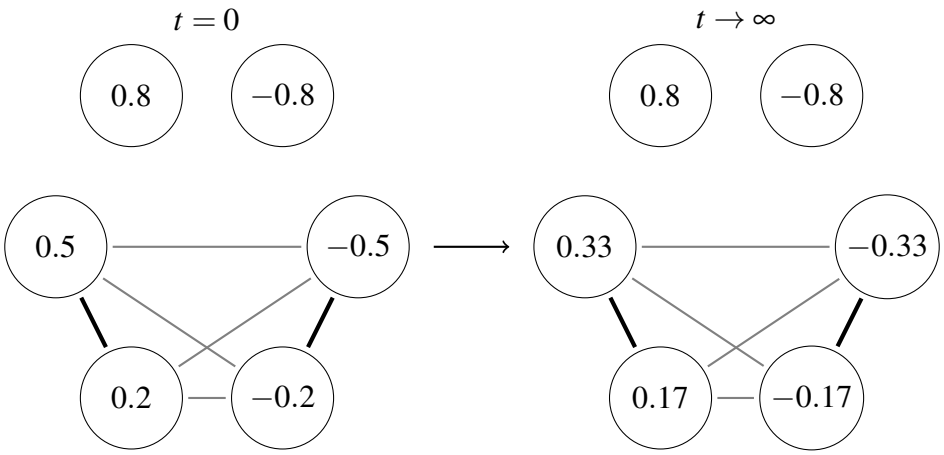

Figure 4: Network with extreme agents banned for $t = 0$ and $t \to \infty$

As we can see the opinions of the extreme agents did not evolve as they were banned, and the other opinions did not changed much, compared to the previous case. The following table sums up the indexes for the both cases:

	non-censored	censored
$\mathcal{P}$	0.66	1.56
$\mathcal{D}$	0.44	0.12
$\mathcal{IC}$	0.32	0.06
$\mathcal{W}$	-1.42	-1.74

Table 1: Indices for non-censored and censored case

As we can see, polarization has increased since the extreme agents did not change their opinions. Disagreement has decreased because the extreme agents did not participate in the discussion. Internal conflict also decreased, as opinions evolved less. The overall welfare[5] decreased, which can be attributed to the choice of α and β. If α were sufficiently low, the welfare would actually increase. Thus, there is a clear trade-off: censorship introduces two opposing forces that impact welfare.

[5] Welfare defined for $\alpha = \beta = \frac{1}{3}$.

3.1 General Case

With all components defined, we can now investigate how the indices change with varying levels of the censor point c. In order to do so, we need to compute the expected values of polarization, disagreement, and internal conflict with respect to distribution of innate opinions, and network formation probabilities.

$$\mathbb{E}[\mathcal{P}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{P}}(\mathbf{L})\bar{\mathbf{s}}] \tag{3.1}$$

$$\mathbb{E}[\mathcal{D}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{D}}(\mathbf{L})\bar{\mathbf{s}}] \tag{3.2}$$

$$\mathbb{E}[\mathcal{IC}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{IC}}(\mathbf{L})\bar{\mathbf{s}}] \tag{3.3}$$

Lemma 1. *The matrices $f_{\mathcal{P}}(\mathbf{L})$, $f_{\mathcal{P}}(\mathbf{D})$ and $f_{\mathcal{P}}(\mathbf{IC})$ can be eigenvalue decomposed into respectively $\mathbf{U}\Lambda_{\mathcal{P}}\mathbf{U}^{\mathbf{T}}$, $\mathbf{U}\Lambda_{\mathcal{D}}\mathbf{U}^{\mathbf{T}}$, and $\mathbf{U}\Lambda_{\mathcal{IC}}\mathbf{U}^{\mathbf{T}}$, where Λ is a diagonal matrix of eigenvalues λ.*

Proof. See appendix □

As shown in Chen, Lijffijt, and De Bie (2018) eigenvalues of $f_{\mathcal{P}}(\mathbf{L})$, $f_{\mathcal{P}}(\mathbf{D})$ and $f_{\mathcal{P}}(\mathbf{IC})$ can be mapped from the eigenvalues of $\mathbf{L}$. For a given eigenvalue of $\mathbf{L}$ λ_L, the eigenvalue of above functions are:

1. $\lambda_{\mathcal{P}} = \frac{1}{(1+\lambda_L)^2}$
2. $\lambda_{\mathcal{D}} = \frac{\lambda_L}{(1+\lambda_L)^2}$
3. $\lambda_{\mathcal{IC}} = \frac{\lambda_L^2}{(1+\lambda_L)^2}$

Given this, each index can be written as:

$$\mathbb{E}[\mathcal{P}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{P}}(\mathbf{L})\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{U}\Lambda_{\mathcal{P}}\mathbf{U}^{\mathbf{T}}\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}_{\mathbf{U}}^{\mathbf{T}}\Lambda_{\mathcal{P}}\bar{\mathbf{s}}_{\mathbf{U}}] = \sum_i^n \lambda_{\mathcal{P}}\mathbb{E}[\bar{s}_{Ui}^2] = \sum_i^n \frac{1}{(1+\lambda_{Li})^2}\mathbb{E}[\bar{s}_{Ui}^2] \tag{3.4}$$

$$\mathbb{E}[\mathcal{D}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{D}}(\mathbf{L})\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{U}\Lambda_{\mathcal{D}}\mathbf{U}^{\mathbf{T}}\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}_{\mathbf{U}}^{\mathbf{T}}\Lambda_{\mathcal{D}}\bar{\mathbf{s}}_{\mathbf{U}}] = \sum_i^n \lambda_{\mathcal{D}}\mathbb{E}[\bar{s}_{Ui}^2] = \sum_i^n \frac{\lambda_{Li}}{(1+\lambda_{Li})^2}\mathbb{E}[\bar{s}_{Ui}^2] \tag{3.5}$$

$$\mathbb{E}[\mathcal{IC}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathcal{IC}}(\mathbf{L})\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{U}\Lambda_{\mathcal{IC}}\mathbf{U}^{\mathbf{T}}\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}_{\mathbf{U}}^{\mathbf{T}}\Lambda_{\mathcal{IC}}\bar{\mathbf{s}}_{\mathbf{U}}] = \sum_i^n \lambda_{\mathcal{IC}}\mathbb{E}[\bar{s}_{Ui}^2] = \sum_i^n \frac{\lambda_{Li}^2}{(1+\lambda_{Li})^2}\mathbb{E}[\bar{s}_{Ui}^2] \tag{3.6}$$

where $\bar{\mathbf{s}}_{\mathbf{U}}$ is a mapping of $\bar{\mathbf{s}}$ onto $\mathbf{U}$.

After finding the eigenvalues[6] λ_{Li} and computing the expectations[7] $\mathbb{E}[\bar{s}_{Ui}^2]$, each index can be expressed as a function of c as follows:

$$\mathbb{E}[\mathcal{P}] = \sigma^2 + \frac{1}{(1+F(c)nq)^2}(n\mu^2+\sigma^2) + (n-2)\frac{1}{(1+F(c)n^{\frac{p+q}{2}})^2}\sigma^2$$

$$\mathbb{E}[\mathcal{D}] = \frac{F(c)nq}{(1+F(c)nq)^2}(n\mu^2+\sigma^2) + (n-2)\frac{F(c)n^{\frac{p+q}{2}}}{(1+F(c)n^{\frac{p+q}{2}})^2}\sigma^2$$

$$\mathbb{E}[\mathcal{IC}] = \frac{(F(c)nq)^2}{(1+F(c)nq)^2}(n\mu^2+\sigma^2) + (n-2)\frac{(F(c)n^{\frac{p+q}{2}})^2}{(1+F(c)n^{\frac{p+q}{2}})^2}\sigma^2$$

Proposition 1. *In the general case:*

1. *Polarization is decreasing in c, thus the higher the censorship, the higher the polarization*
2. *Internal conflict is increasing in c, thus the higher the censorship, the lower the internal conflict*

Proof. See appendix □

This result suggests that if the social media platform increases the level of censorship, the society is expected to be become more polarized in equilibrium. This increased polarization may occur because banned agents are unable to communicate, leaving their opinions unchanged and

[6] See appendix for detailed derivations.

[7] See appendix for detailed derivations.

uninfluenced by others. Consequently, these opinions remain more distant compared to a scenario with lower censorship.

On the other hand, as censorship increases, the society is expected to experience less internal conflict in equilibrium. This may be due to the same reason: with higher censorship, agents communicate less, resulting in fewer changes in opinions and thus reduced internal conflict.

When it comes to disagreement, predicting its behavior with increasing censorship is extremely complex due to algebraic intricacies. To better understand the effects of censorship on disagreement, as well as on polarization and internal conflict, we have plotted[8] these indices, depending on c, for uniform distribution, $n = 100$, and different sets of parameters, p, and q for illustrative purposes.

[8]The plots of polarization, disagreement, internal conflict, and welfare with interactive parameters n, p, q, α, and β can be seen at https://mcooocm.github.io/.

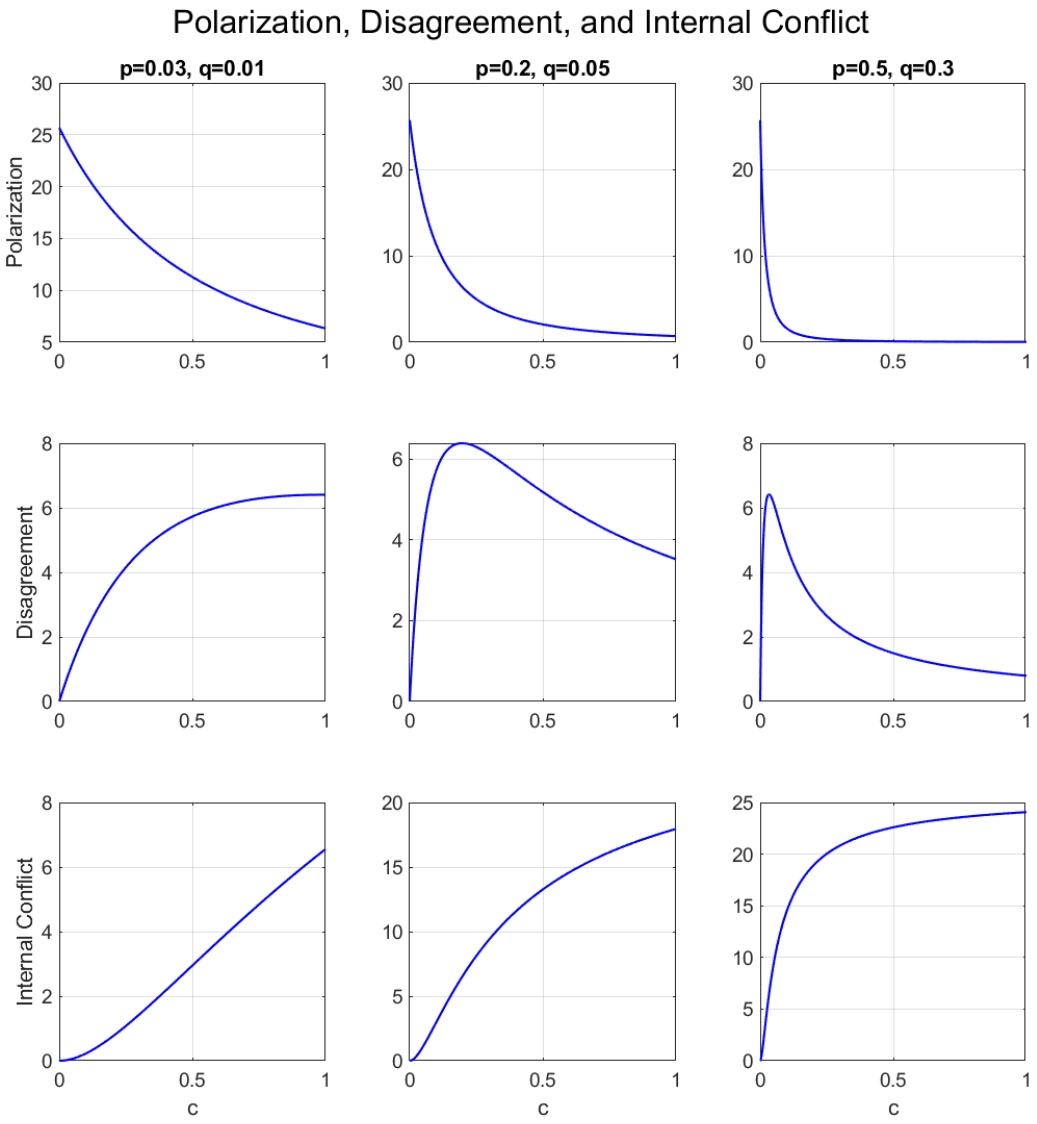

Figure 5: Polarization, Disagreement and Internal Conflict depending on c, for uniform distribution, $n = 100$, and different sets of p, q

In the figure, it is evident that initially, the disagreement is zero, but then it increases. For the last two columns, the disagreement reaches a maximum at some point before decreasing. Additionally, it can be observed that polarization decreases with increasing c in each case, while internal conflict increases with c. Whatsmore, the values of p and q increase, polarization diminishes more rapidly, while internal conflict escalates more quickly.

Due to the algebraic complexity, finding this optimal level requires numerical methods for given parameters. The plots below present the welfare and optimal levels of censorship for uniform distribution and different parameters values p, q ,α, and β.

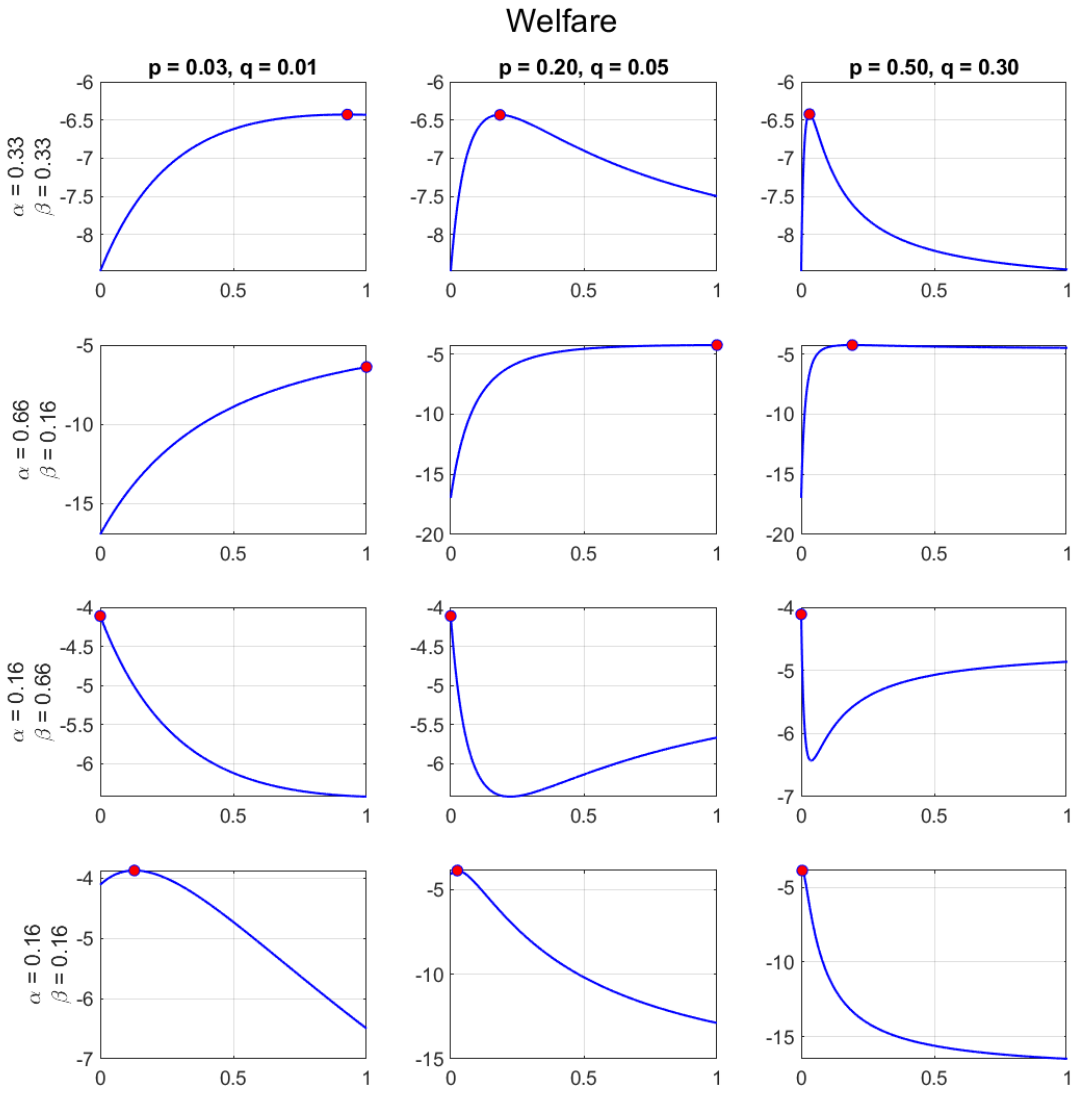

Figure 6: Welfare depending on c, for uniform distribution, $n = 100$, and different sets of p, q, α, β

Depending on the parameter values, it might be optimal to ban everyone, no one, or just part of society. When all indexes are treated equally, the optimal censorship level largely depends on

the network's connectedness. If the network administrator prioritizes reducing polarization[9], it is optimal not to ban anyone, as polarization tends to increase with censorship. When the focus is on minimizing disagreement[10], banning everyone is optimal, as this results in zero disagreement. Lastly, when internal conflict is the most relevant index, the network administrator should decide to ban all or almost all agents, as internal conflict decreases withh censorship.

3.2 Extremely High Homophily

Extremely high homophily would mean that agents only communicate within their own group. In the model, this is represented by setting the parameter $q = 0$, as q indicates the intensity of communication across groups, and with it set to 0, there is no cross-communication. In this highly homophilic scenario, the indexes would become much more simplified, specifically:

$$\mathbb{E}[\mathcal{P}] = n\mu^2 + 2\sigma^2 + (n-2)\frac{1}{(1+F(c)n\frac{p}{2})^2}\sigma^2$$

$$\mathbb{E}[\mathcal{D}] = (n-2)\frac{F(c)n\frac{p}{2}}{(1+F(c)n\frac{p}{2})^2}\sigma^2$$

$$\mathbb{E}[\mathcal{IC}] = (n-2)\frac{(F(c)n\frac{p}{2})^2}{(1+F(c)n\frac{p}{2})^2}\sigma^2$$

Proposition 2. *In the highly homophilic case:*

1. *Polarization is decreasing in c, thus the higher the censorship, the higher the polarization*
2. *Internal conflict is increasing in c, thus the higher the censorship, the lower the internal conflict*

Proof. It follows from proof of Proposition 1, with setting $q = 0$ □

It follows the same logic as in previous case.

[9] High α.
[10] High β.

Proposition 3. *In the highly homophilic case the disagreement is:*

1. *increasing in c if* $\frac{2}{np} > 1$
2. *increasing in* $c \in [0, Q(\frac{2}{np}))$*, and decreasing in* $c \in (Q(\frac{2}{np}), 1]$ *if* $\frac{2}{np} < 1$

,where $Q()$ *is a quantile function of* $F()$*.*

Proof. See appendix □

This result indicates that if the number of agents and the internal connectivity are sufficiently high, there exists a threshold c below which disagreement increases and above which it decreases. This implies that initially, as censorship increases, disagreement will also increase. Thus, at first, the effect of reduced communication is stronger than the diminishing connectivity between agents. However, beyond a certain point, the impact of reduced connectivity becomes more significant. Polarization, disagreement and internal conflict with varying levels of c are plotted below for uniform distribution, $n = 100$, and varying level of p.

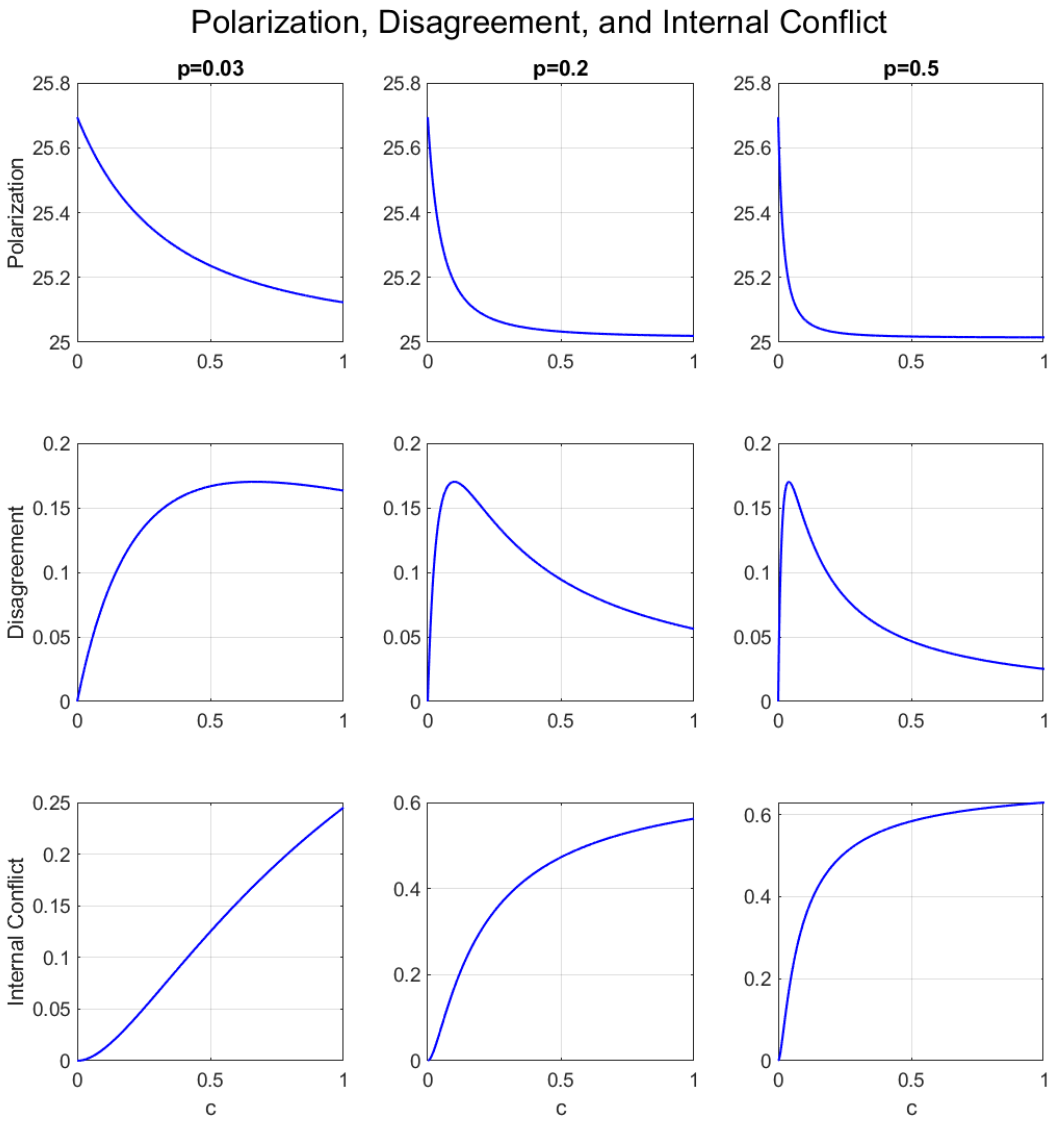

Figure 7: Polarization, Disagreement and Internal Conflict depending on c, for uniform distribution, $n = 100$, and different sets of p

Interestingly, compared to the general case, the indexes show minimal change. This is because, in the high homophily scenario, agents predominantly communicate with like-minded individuals, resulting in opinions that remain close to their innate ones. The plots indicate that for $\frac{2}{np}$[11], disagreement starts to decrease.

[11]There is no quantile function as the distribution is uniform on $[0,1]$.

Proposition 4. *In the highly homophilic case, the c that maximizes the welfare is:*

1. $c^* = 1$, *if* $\alpha > \frac{\beta(1-3n\frac{p}{2})+np}{2(1+n\frac{p}{2})}$ *and* $\alpha > \frac{\beta(1-n\frac{p}{2})+n\frac{p}{2}}{2(1+n\frac{p}{2})}$
2. $c^* = Q(\frac{2}{np}\frac{2\alpha-\beta}{2-2\alpha-3\beta})$, *if* $\alpha > \frac{\beta}{2}$ *and* $\alpha < \frac{\beta(1-3n\frac{p}{2})+np}{2(1+n\frac{p}{2})}$
3. $c^* = 0$, *if* $\alpha < \frac{\beta}{2}$ *and* $\alpha < \frac{\beta(1-n\frac{p}{2})+n\frac{p}{2}}{2(1+n\frac{p}{2})}$

,where $Q()$ *is a quantile function of* $F()$.

Proof. See appendix □

Which can be seen more intuitively in the following figure:

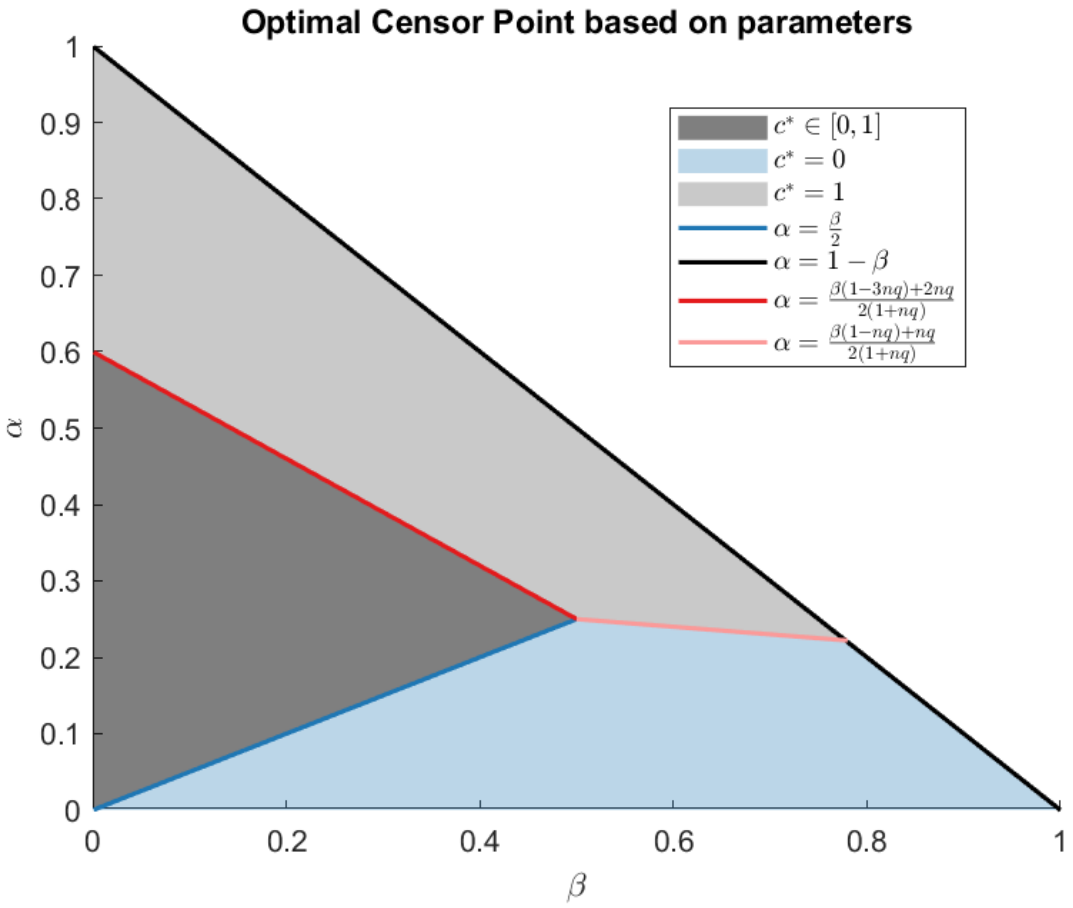

Figure 8: Optimal censor point based on α, and β for $n = 100$, and $p = 0.015$

If the weight of polarization is too high, it will always be optimal to censor everyone. Conversely, if insufficient weight is placed on polarization, it will always be optimal not to censor

anyone. Additionally, as n or p increases, the red line will approach $\alpha = 1 - \frac{3}{2}\beta$, while the pink line will approach $\alpha = \frac{1-\beta}{2}$. The limiting case, where $nq \to \infty$ can be seen in the next figure.

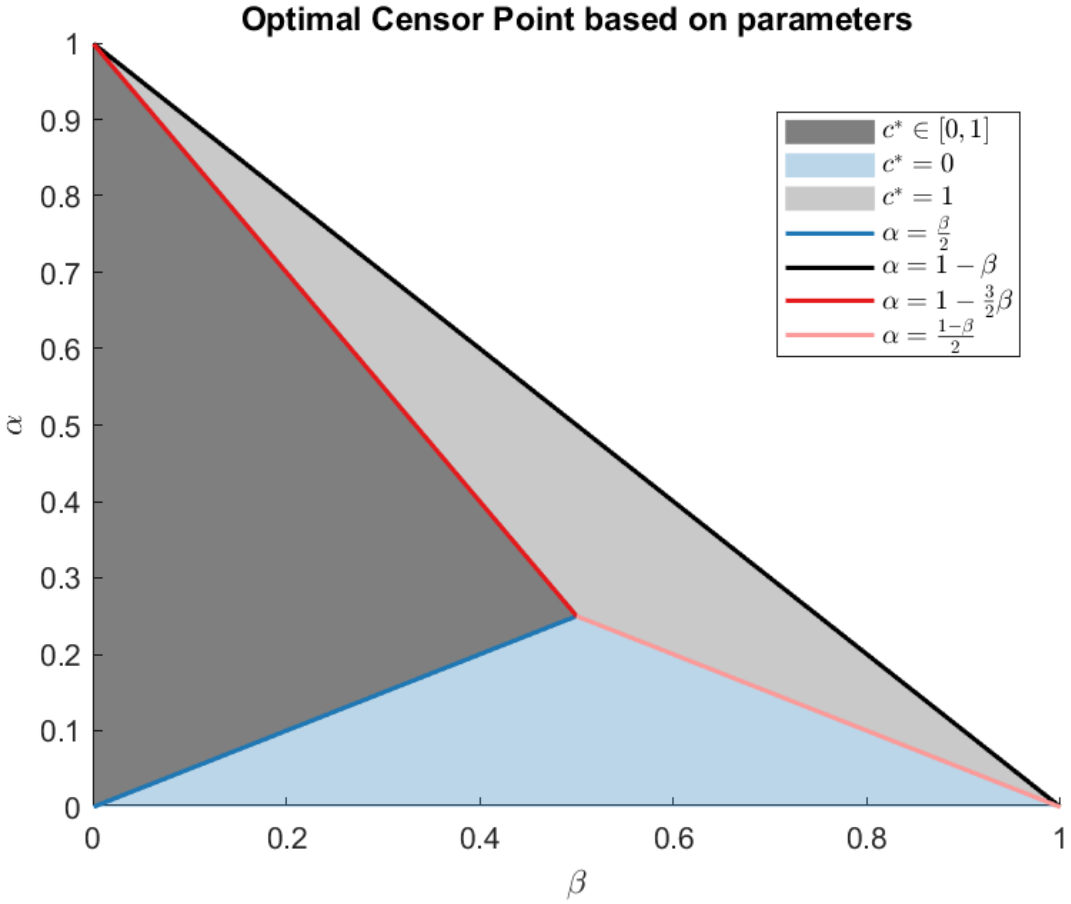

Figure 9: Optimal censor point based on α, and β for $np \to \infty$

This indicates that the parameter space in which the optimal c^* lies within $[0,1]$ expands with increasing n and p, while the parameter space for which the optimal $c^* = 0$ decreases as n and p increase. The next figure presents Welfare values for different parameter values.

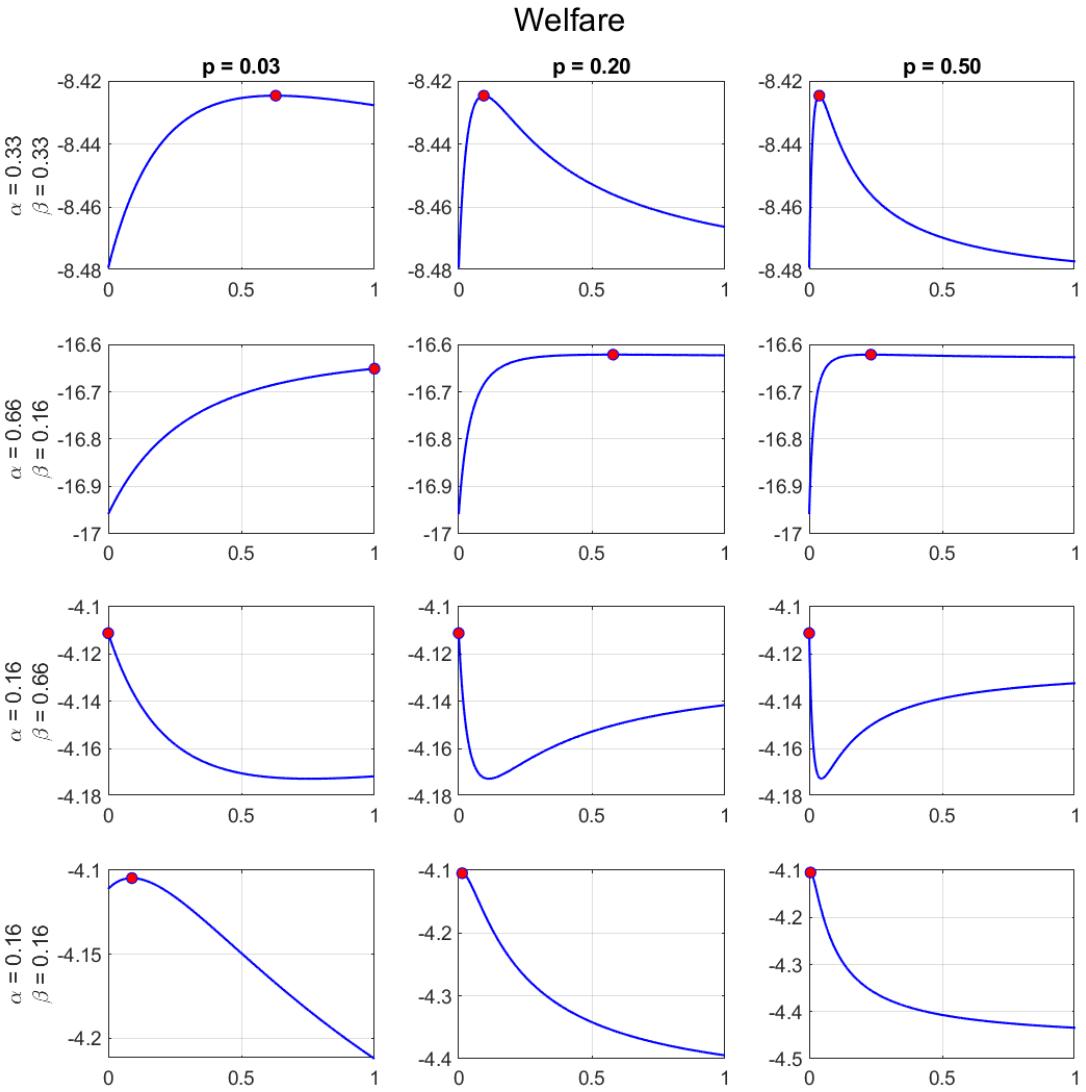

Figure 10: Welfare depending on c, for uniform distribution, $n = 100$, and different sets of p, q, α, β

Here, it can also be observed that the changes are minimal compared to the general case, while the same logic still applies, regarding the optimal censor point.

3.3 Extremely Low Homophily

Extremely low homophily would mean that agents communicate with the other groups as much as with their own. As q indicates the intensity of communication across groups, and p indicates the intensity of communication within the groups, then for the low homophily it is required $q = p$. In this extremely low homophilic scenario, the indexes would become much more simplified, specifically:

$$\mathbb{E}[\mathcal{P}] = \sigma^2 + \frac{1}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathcal{D}] = \frac{F(c)np}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathcal{IC}] = \frac{(F(c)np)^2}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

Proposition 5. *In the extremely low homophilic case:*

1. *Polarization is decreasing in c, thus the higher the censorship, the higher the polarization*
2. *Internal conflict is increasing in c, thus the higher the censorship, the lower the internal conflict*

Proof. It follows from proof of Proposition 1, with setting $q = p$ □

Proposition 6. *In the highly homophilic case the disagreement is:*

1. *increasing in c if* $\frac{1}{np} > 1$
2. *increasing in* $c \in [0, Q(\frac{1}{np}))$*, and decreasing in* $c \in (Q(\frac{1}{np}), 1]$ *if* $\frac{1}{np} < 1$

,where $Q()$ is a quantile function of $F()$.

Proof. Follows from proof of Proposition 3 with setting $q = p$ □

It is important to note that the results are very similar to those in the previous section. Specifically, there exists a threshold c below which disagreement decreases and above which it increases, occurring two times slower than in the case of extremely high homophily. Networks with high

connections across groups tend to have a more intense response to censorship. The plots below illustrate polarization, disagreement, and internal conflict for varying levels of c, assuming a uniform distribution, $n = 100$, and varying levels of p.

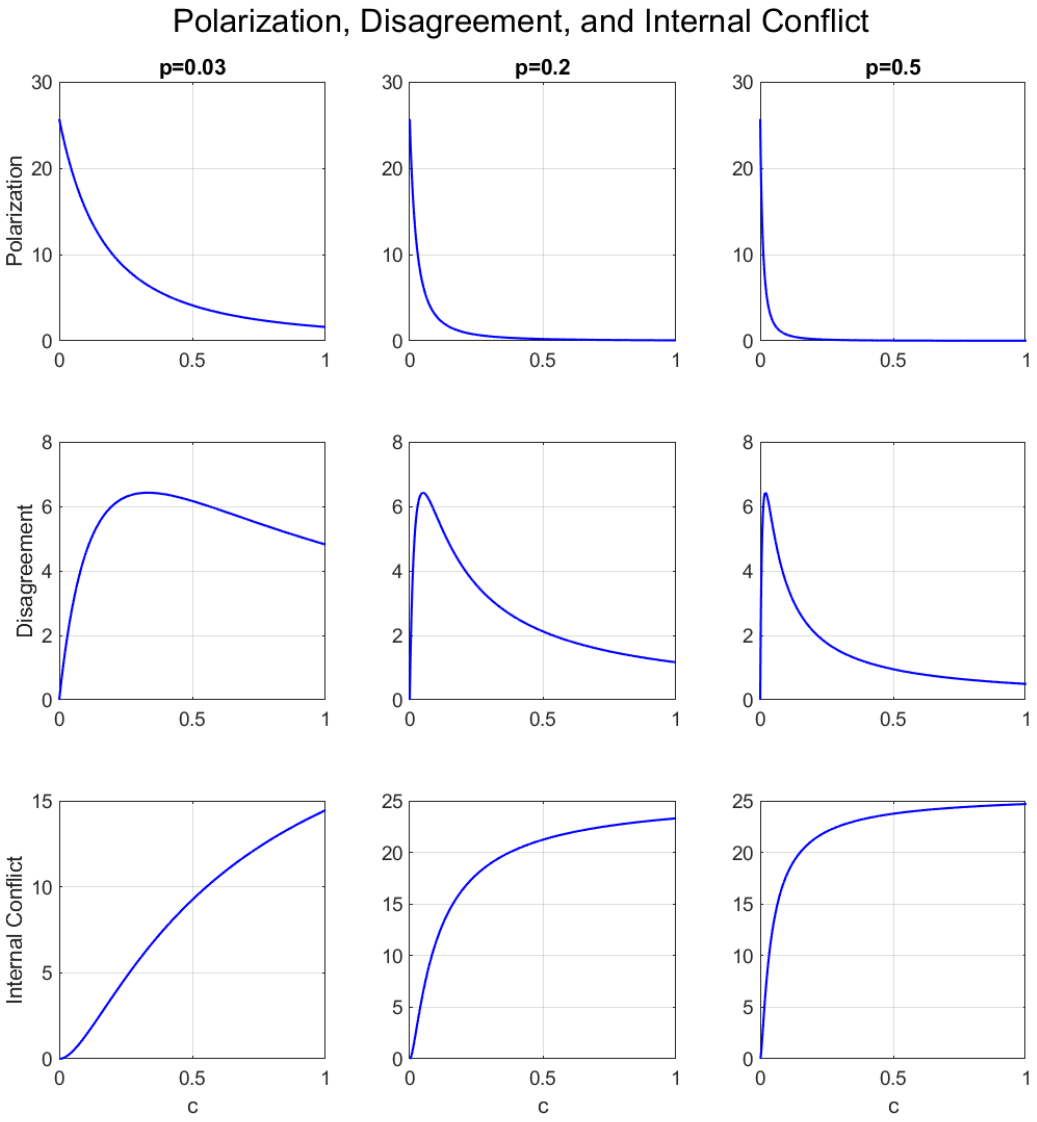

Figure 11: Polarization, Disagreement and Internal Conflict depending on c, for uniform distribution, $n = 100$, and different sets of p

Proposition 7. *In the extremely low homophilic case, the c that maximizes the Welfare is:*

1. $c^* = 1$, *if* $\alpha > \frac{\beta(1-3np)+2np}{2(1+np)}$ *and* $\alpha > \frac{\beta(1-np)+np}{2(1+np)}$
2. $c^* = Q(\frac{1}{np}\frac{2\alpha-\beta}{2-2\alpha-3\beta})$, *if* $\alpha > \frac{\beta}{2}$ *and* $\alpha < \frac{\beta(1-3np)+2np}{2(1+np)}$
3. $c^* = 0$, *if* $\alpha < \frac{\beta}{2}$ *and* $\alpha < \frac{\beta(1-np)+np}{2(1+np)}$

,where $Q()$ *is a quantile function of* $F()$.

Proof. Follows from proof of Proposition 4 with setting $q = p$ □

It can be seen that the results do not vary much, than in the previous section. Here again the difference is that the optimal censorship would be higher, and the parameter space for which the optimal censor $c^* \in [0,1]$ is higher than in the high homophily case. In the $np \to \infty$ the cases are identical. Next figure presents Welfare values for different parameter values.

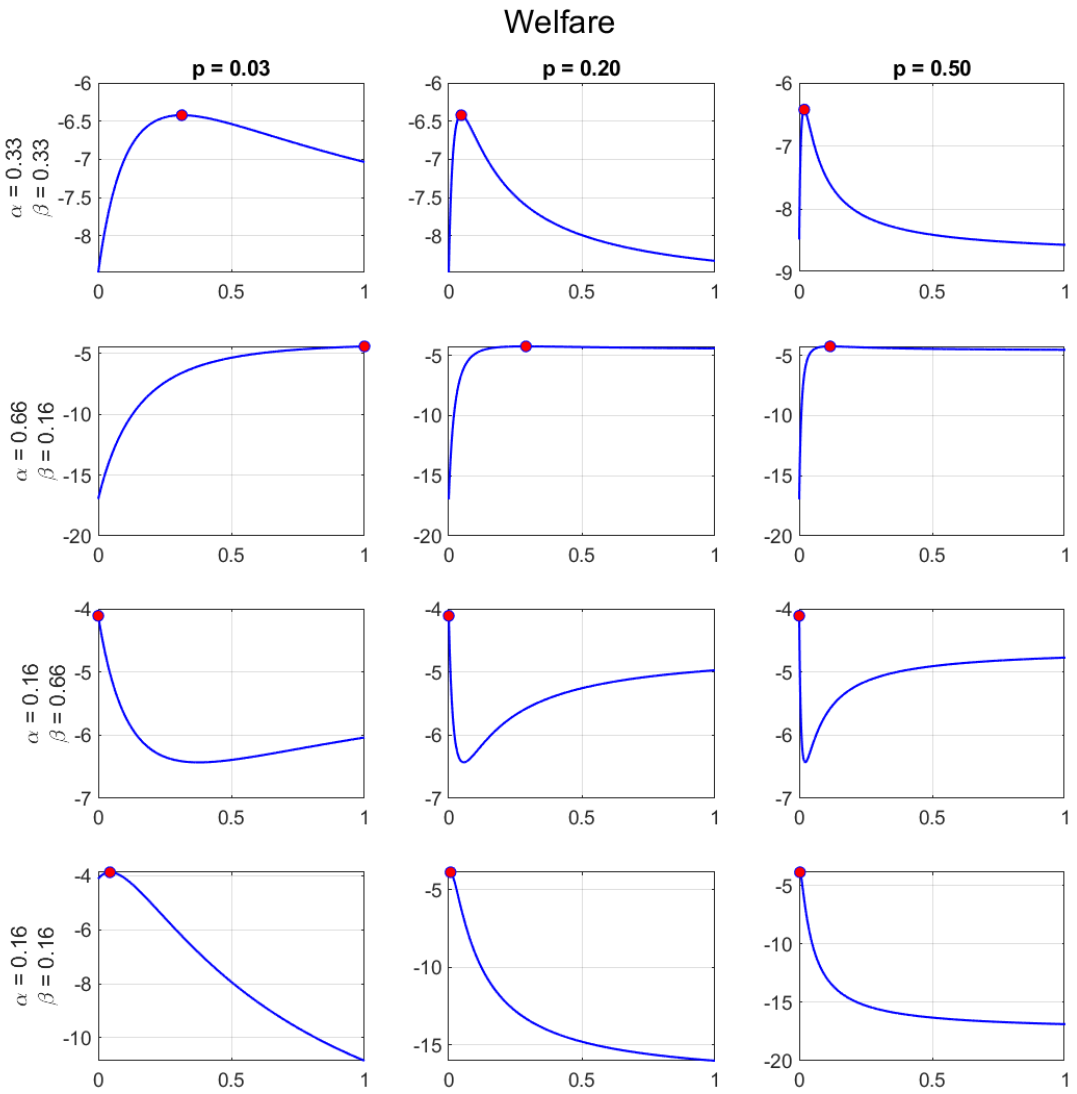

Figure 12: Welfare depending on c, for uniform distribution, $n = 100$, and different sets of p, q, α, β

The differences between the low and high homophily cases can be intuitively explained. When opposing groups communicate with each other, it leads to an equilibrium where opinions are closer together compared to the case where opposing groups do not communicate. As a result, censorship has a more significant impact on the indices in scenarios where agents can communicate across groups, as it severely restricts the potential for opinions to evolve.

4 Conclusions

To the best of our knowledge, this is the first paper to investigate the effects of censorship in social media. We employed the Friedkin-Johansen Opinion Dynamics Model and assumed two types of agents with opposing opinion distributions. Censorship is introduced by assuming the existence of a network administrator who sets a threshold; agents with opinions outside this threshold are banned. We examined the effects of censorship in terms of polarization, disagreement, internal conflict, and a weighted sum of these factors defined as welfare from the perspective of the network administrator.

Our results show that, generally, higher levels of censorship lead to increased polarization and reduced internal conflict in society. Disagreement initially increases with censorship but decreases to zero after a certain point. Additionally, for specific homophily scenarios, we identified parameter spaces where an optimal censorship threshold exists. The results also indicate that the effects of censorship are higher in networks where agents communicate more across groups. As the number of agents approaches infinity, the cases with extremely high and extremely low homophily converge.

Future research should focus on investigating the incentives of the network administrator. For social media platform owners, the goal is not welfare maximization but profit maximization. This could involve incorporating the costs of censorship and defining a profit function based on user activity. Furthermore, it would be interesting to explore the empirical counterpart of the model by estimating parameters using datasets from platforms like Twitter or Reddit.

References

Acemoglu, Daron and Asuman Ozdaglar (2011). "Opinion dynamics and learning in social networks". In: *Dynamic Games and Applications* 1, pp. 3–49.

Alkiviadou, Natalie (2019). "Hate speech on social media networks: towards a regulatory framework?" In: *Information & Communications Technology Law* 28(1), pp. 19–35.

Allcott, Hunt et al. (2020). "The welfare effects of social media". In: *American Economic Review* 110(3), pp. 629–676.

Baer, Drake (2016). "The "filter bubble" explains why trump won and you didn't see it coming". In: *New York Magazine*.

Bao, Te, Bin Liang, and Yohanes E Riyanto (2021). "Unpacking the negative welfare effect of social media: Evidence from a large scale nationally representative time-use survey in China". In: *China Economic Review* 69, p. 101650.

Braha, Dan and Marcus AM De Aguiar (2017). "Voting contagion: Modeling and analysis of a century of US presidential elections". In: *PloS one* 12(5), e0177970.

Breyette, Sarah K and Katharine Hill (2015). "The impact of electronic communication and social media on child welfare practice". In: *Journal of Technology in Human Services* 33(4), pp. 283–303.

Chen, Xi, Jefrey Lijffijt, and Tijl De Bie (2018). "Quantifying and minimizing risk of conflict in social networks". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1197–1205.

Chitra, Uthsav and Christopher Musco (2020). "Analyzing the impact of filter bubbles on social network polarization". In: *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 115–123.

Clifford, Peter and Aidan Sudbury (1973). "A model for spatial conflict". In: *Biometrika* 60(3), pp. 581–588.

Deffuant, Guillaume et al. (2000). "Mixing beliefs among interacting agents". In: *Advances in Complex Systems* 3(01n04), pp. 87–98.

DeGroot, Morris H (1974). "Reaching a consensus". In: *Journal of the American Statistical association* 69(345), pp. 118–121.

Dehghan, Ehsan (2018). "A year of discursive struggle over freedom of speech on Twitter: What can a mixed-methods approach tell us?" In: *Proceedings of the 9th International Conference on Social Media and Society*, pp. 266–270.

Dixon, SJ (2023). *X/twitter: Number of worldwide users 2019-2024. Statista*.

Friedkin, Noah E and Eugene C Johnsen (1997). "Social positions in influence networks". In: *Social networks* 19(3), pp. 209–222.

Holone, Harald (2016). "The filter bubble and its effect on online personal health information". In: *Croatian medical journal* 57(3), p. 298.

McCright, Aaron M and Riley E Dunlap (2011). "The politicization of climate change and polarization in the American public's views of global warming, 2001–2010". In: *The Sociological Quarterly* 52(2), pp. 155–194.

Musco, Cameron, Christopher Musco, and Charalampos E Tsourakakis (2018). "Minimizing polarization and disagreement in social networks". In: *Proceedings of the 2018 world wide web conference*, pp. 369–378.

Ortiz-Ospina, Esteban (2019). "The rise of social media". In: *Our World in Data*. https://ourworldindata.org/rise-of-social-media.

Quattrociocchi, Walter, Guido Caldarelli, and Antonio Scala (2014). "Opinion dynamics on interacting networks: media competition and social influence". In: *Scientific reports* 4(1), p. 4938.

Sage, Melanie and Todd Sage (2016). "Social media and e-professionalism in child welfare: Policy and practice". In: *Journal of Public Child Welfare* 10(1), pp. 79–95.

Segado-Boj, Francisco and Jesús Díaz del Campo Lozano (2020). "Social media and its intersections with free speech, freedom of information and privacy: an analysis". In: *Icono14* 18(1), pp. 231–255.

Smith, Aaron and Monica Anderson (2018). "Social media use in 2018". In.

Smith, Kirsten P and Nicholas A Christakis (2008). "Social networks and health". In: *Annu. Rev. Sociol* 34, pp. 405–429.

Xia, Haoxiang, Huili Wang, and Zhaoguo Xuan (2011). "Opinion dynamics: A multidisciplinary review and perspective on future research". In: *International Journal of Knowledge and Systems Science (IJKSS)* 2(4), pp. 72–91.

A Appendix

A.1 Derivations for Internal Conflict

$$\sum_i^n (\bar{z}_i - \bar{s}_i)^2 = \sum_i^n (\bar{z}_i^2 - 2\bar{z}_i\bar{s}_i + \bar{s}_i^2) = \bar{\mathbf{z}}^\mathbf{T}\bar{\mathbf{z}} - 2\bar{\mathbf{z}}^\mathbf{T}\bar{\mathbf{s}} + \bar{\mathbf{s}}^\mathbf{T}\bar{\mathbf{s}} = \bar{\mathbf{s}}^\mathbf{T}(\mathbf{I}+\mathbf{L})^{-\mathbf{2}}\bar{\mathbf{s}} - 2\bar{\mathbf{s}}^\mathbf{T}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\bar{\mathbf{s}} + \bar{\mathbf{s}}^\mathbf{T}\bar{\mathbf{s}} =$$
$$= \bar{\mathbf{s}}^\mathbf{T}[(\mathbf{I}+\mathbf{L})^{-\mathbf{2}} - 2(\mathbf{I}+\mathbf{L})^{-\mathbf{1}} + \mathbf{I}]\bar{\mathbf{s}} = \bar{\mathbf{s}}^\mathbf{T}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}[\mathbf{I} - \mathbf{2}(\mathbf{I}+\mathbf{L}) + (\mathbf{I}+\mathbf{L})^{\mathbf{2}}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\bar{\mathbf{s}}$$
$$= \bar{\mathbf{s}}^\mathbf{T}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{L}^{\mathbf{2}}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\bar{\mathbf{s}}$$

A.2 Proof of Lemma 1

Each of these matrices are square and symmetric, with the same linearly independent eigenvectors that are also eigenvectors of $\mathbf{L}$. Here it can be seen that L has the same eigenvectors as $f_\mathcal{P}(\mathbf{L})$:

$$\mathbf{L}\mathbf{v} = \lambda_L\mathbf{v} \iff (\mathbf{I}+\mathbf{L})\mathbf{v} = (1+\lambda_L)\mathbf{v} \iff (\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{1}{1+\lambda_L}\mathbf{v} \iff$$
$$(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{1}{1+\lambda_L}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} \iff (\mathbf{I}+\mathbf{L})^{-\mathbf{2}}\mathbf{v} = \frac{1}{(1+\lambda_L)^2}\mathbf{v} \iff$$
$$f_\mathcal{P}(\mathbf{L})\mathbf{v} = \lambda_\mathcal{P}\mathbf{v}$$

Here it can be seen that L has the same eigenvectors as $f_\mathcal{D}$:

$$\mathbf{L}\mathbf{v} = \lambda_L\mathbf{v} \iff (\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{1}{1+\lambda_L}\mathbf{v} \iff \mathbf{L}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{1}{1+\lambda_L}\mathbf{L}\mathbf{v} \iff$$
$$\mathbf{L}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{\lambda_L}{1+\lambda_L}\mathbf{v} \iff (\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{L}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{\lambda_L}{1+\lambda_L}(I+L)^{-1}\mathbf{v} \iff$$
$$(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{L}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{\lambda_L}{(1+\lambda_L)^2}\mathbf{v} \iff$$
$$f_\mathcal{D}(\mathbf{L})\mathbf{v} = \lambda_\mathcal{D}\mathbf{v}$$

Here it can be seen that L has the same eigenvectors as $f_\mathcal{JC}$:

$$\mathbf{L}\mathbf{v} = \lambda_L\mathbf{v} \iff (\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{1}{1+\lambda_L}\mathbf{v} \iff \mathbf{L}^{\mathbf{2}}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{1}{1+\lambda_L}\mathbf{L}^{\mathbf{2}}\mathbf{v} \iff$$
$$\mathbf{L}^{\mathbf{2}}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{\lambda_L^2}{1+\lambda_L}\mathbf{v} \iff (\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{L}^{\mathbf{2}}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{\lambda_L^2}{1+\lambda_L}(I+L)^{-1}\mathbf{v} \iff$$
$$(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{L}^{\mathbf{2}}(\mathbf{I}+\mathbf{L})^{-\mathbf{1}}\mathbf{v} = \frac{\lambda_L^2}{(1+\lambda_L)^2}\mathbf{v} \iff$$
$$f_\mathcal{JC}(\mathbf{L})\mathbf{v} = \lambda_\mathcal{JC}\mathbf{v}$$

A.3 Derivations of λ_{Li}

It is common knowledge in the SBM literature that the eigenvalues of **A** are:

$$\lambda_{A1} = F(c)n\frac{p+q}{2}, \lambda_{A2} = F(c)n\frac{p-q}{2}, \lambda_{A3} = ...\lambda_{An} = 0$$

And the eigenvalues of **D** are:

$$\lambda_{D1} = \lambda_{D2} = ...\lambda_{Dn} = F(c)n\frac{p+q}{2}$$

, then the eigenvalues of **L** are:

$$\lambda_{L1} = 0, \lambda_{L2} = F(c)qn, \lambda_{L3} = ... = \lambda_{Ln} = F(c)n\frac{p+q}{2}$$

And from Chen, Lijffijt, and De Bie (2018) we know the mappings from eigenvalues of **L** to eigenvalues of each index:

1. $\lambda_{\mathcal{P}} = \frac{1}{(1+\lambda_L)^2}$
2. $\lambda_{\mathcal{D}} = \frac{\lambda_L}{(1+\lambda_L)^2}$
3. $\lambda_{\mathcal{JC}} = \frac{\lambda_L^2}{(1+\lambda_L)^2}$

,thus:

$$\lambda_{\mathcal{P}1} = 1, \lambda_{\mathcal{P}2} = \frac{1}{(1+F(c)qn)^2}, \lambda_{\mathcal{P}3} = ... = \lambda_{\mathcal{P}n} = \frac{1}{(1+F(c)\frac{p+q}{2}n)^2},$$

$$\lambda_{\mathcal{D}1} = 0, \lambda_{\mathcal{D}2} = \frac{F(c)qn}{(1+F(c)qn)^2}, \lambda_{\mathcal{D}3} = ... = \lambda_{\mathcal{D}n} = \frac{F(c)\frac{p+q}{2}n}{(1+F(c)\frac{p+q}{2}n)^2},$$

$$\lambda_{\mathcal{JC}1} = 0, \lambda_{\mathcal{JC}2} = \frac{(F(c)qn)^2}{(1+F(c)qn)^2}, \lambda_{\mathcal{JC}3} = ... = \lambda_{\mathcal{JC}n} = \frac{(F(c)\frac{p+q}{2}n)^2}{(1+F(c)\frac{p+q}{2}n)^2},$$

A.4 Derivations of $\mathbb{E}[\bar{s}_{Ui}^2]$

Note that the eigenvectors of **A** and **L** are the same, thus as the eigenvectors of **L** and each index are also the same, then **U** is a matrix of orthonormalized eigenvectors of **A**, thus:

1. First eigenvalue is $\lambda_1 = n\frac{p+q}{2}$ with eigenvector $\mathbf{v_1} = \frac{\vec{1}}{\sqrt{n}}$ (to ensure it is unit)

2. Second eigenvalue is $\lambda_2 = n\frac{p-q}{2}$ with eigenvector $v_{2i} = \begin{cases} \frac{1}{\sqrt{n}} & if \quad i \leq n/2 \\ \frac{-1}{\sqrt{n}} & if \quad i > n/2 \end{cases}$

3. Since $rankA = 2$ rest of eigenvalues are 0 and eigenvectors v are such that $Av = 0$, are unit vectors and are orthogonal to each other. $Av = 0$ means that both elements on $1 : \frac{n}{2}$ and $\frac{n}{2} + 1 : n$ have to sum up to 0. Thus let me propose that the rest of the eigenvectors take the following structure:

   - if $k \in (3, 4, ..., n/2+1)$, then:

$$\mathbf{v_k^T} = [\underbrace{\underbrace{\frac{1}{\sqrt{(k-1)(k-2)}} \dots \frac{1}{\sqrt{(k-1)(k-2)}} \quad \frac{-(k-2)}{\sqrt{(k-1)(k-2)}}}_{(k-1) \text{ elements}} \quad 0 \dots 0}_{\frac{n}{2} \text{ elements}} \quad \underbrace{0 \dots 0}_{\frac{n}{2} \text{elements}}]$$

   - if $k \in (n/2+2, ..., n)$, then:

$$\mathbf{v_k^T} = [\underbrace{0 \dots 0}_{\frac{n}{2} \text{elements}} \underbrace{\underbrace{\frac{1}{\sqrt{(k-1)(k-2)}} \dots \frac{1}{\sqrt{(k-1)(k-2)}} \quad \frac{-(k-2)}{\sqrt{(k-1)(k-2)}}}_{(k-1) \text{ elements}} \quad 0 \dots 0}_{\frac{n}{2} \text{ elements}} \quad]$$

Thus the whole $\mathbf{U}$ matrix is:

$$\mathbf{U} = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \dots & \frac{1}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}} & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \dots & \frac{1}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}} & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & 0 & \frac{-2}{\sqrt{6}} & \dots & \frac{1}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & 0 & 0 & \dots & \frac{-(\frac{n}{2}-1)}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}} & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{n}} & \frac{-1}{\sqrt{n}} & 0 & 0 & \dots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \dots & \frac{1}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}} \\ \frac{1}{\sqrt{n}} & \frac{-1}{\sqrt{n}} & 0 & 0 & \dots & 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \dots & \frac{1}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}} \\ \frac{1}{\sqrt{n}} & \frac{-1}{\sqrt{n}} & 0 & 0 & \dots & 0 & 0 & \frac{-2}{\sqrt{6}} & \dots & \frac{1}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n}} & \frac{-1}{\sqrt{n}} & 0 & 0 & \dots & 0 & 0 & 0 & \dots & \frac{-(\frac{n}{2}-1)}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}} \end{bmatrix}$$

Each of those vectors is a unit vector, they are independent, and vectors 2 to n add up to 0. Moreover $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. For future references belongs to L/ R will indicate that the agent is distributed according to F_L/F_R

Using this we can compute the projection of $\bar{\mathbf{s}}$ onto $\mathbf{U}$. I will do it one by one:

$$\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{v_1} = [\overbrace{\bar{s}_1 \quad \dots \quad \bar{s}_{\frac{n}{2}}}^{\text{belongs to } L} \quad \overbrace{\bar{s}_{\frac{n}{2}+1} \quad \dots \quad \bar{s}_n}^{\text{belongs to } R}] \begin{bmatrix} \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{bmatrix} = \frac{1}{\sqrt{n}} \sum_i^n \bar{s}_i$$

$$\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{v_2} = [\overbrace{\bar{s}_1 \quad \dots \quad \bar{s}_{\frac{n}{2}}}^{\text{belongs to } L} \quad \overbrace{\bar{s}_{\frac{n}{2}+1} \quad \dots \quad \bar{s}_n}^{\text{belongs to } R}] \begin{bmatrix} \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \\ \frac{-1}{\sqrt{n}} \\ \vdots \\ \frac{-1}{\sqrt{n}} \end{bmatrix} = \frac{1}{\sqrt{n}} (\sum_{i=1}^{\frac{n}{2}} \bar{s}_i - \sum_{i=\frac{n}{2}+1}^{n} \bar{s}_i)$$

for $k \in (3,4,...,n/2+1)$:

$$\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{v_k} = [\overbrace{\bar{s}_1 \quad \dots \quad \bar{s}_{\frac{n}{2}}}^{\text{belongs to } L} \quad \overbrace{\bar{s}_{\frac{n}{2}+1} \quad \dots \quad \bar{s}_n}^{\text{belongs to } R}] \begin{bmatrix} \frac{1}{\sqrt{(k-1)(k-2)}} \\ \vdots \\ \frac{-(k-2)}{\sqrt{(k-1)(k-2)}} \\ \vdots \\ 0 \end{bmatrix} = \frac{1}{\sqrt{(k-1)(k-2)}} (\sum_{i=1}^{k-2} \bar{s}_i - (k-2)\bar{s}_{k-1})$$

It is symmetrical in expectations to the $k \in (n/2+2,...,n)$ case.

Having that we can construct the $\bar{\mathbf{s}}_{\mathbf{U}}^{\mathbf{T}}$:

$$\bar{\mathbf{s}}_{\mathbf{U}}^{\mathbf{T}} = \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_i^n \bar{s}_i \\ \frac{1}{\sqrt{n}}(\sum_{i=1}^{\frac{n}{2}} \bar{s}_i - \sum_{i=\frac{n}{2}+1}^{n} \bar{s}_i) \\ \frac{1}{\sqrt{2}}(\bar{s}_1 - \bar{s}_2) \\ \frac{1}{\sqrt{6}}(\bar{s}_1 + \bar{s}_2 - 2\bar{s}_3) \\ \vdots \\ \frac{1}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}}(\sum_{i=1}^{\frac{n}{2}-1} \bar{s}_i - (\frac{n}{2}-1)\bar{s}_{\frac{n}{2}}) \\ \frac{1}{\sqrt{2}}(\bar{s}_{\frac{n}{2}+1} - \bar{s}_{\frac{n}{2}+2}) \\ \frac{1}{\sqrt{6}}(\bar{s}_{\frac{n}{2}+1} + \bar{s}_{\frac{n}{2}+2} - 2\bar{s}_{\frac{n}{2}+3}) \\ \vdots \\ \frac{1}{\sqrt{\frac{n}{2}(\frac{n}{2}-1)}}(\sum_{i=\frac{n}{2}+1}^{n} \bar{s}_i - (\frac{n}{2}-1)\bar{s}_n) \end{bmatrix}$$

And finally:

$$\begin{aligned} \mathbb{E}[\bar{s}_{U1}^2] &= \mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_i^n \bar{s}_i\right)\left(\frac{1}{\sqrt{n}}\sum_i^n \bar{s}_i\right)\right] = \frac{1}{n}\mathbb{E}\left[\left(\sum_i^n \bar{s}_i\right)^2\right] = \\ &= \frac{1}{n}\mathbb{E}\left[\sum_i^n\sum_j^n \bar{s}_i\bar{s}_j\right] = \frac{1}{n}\left[n(\mu^2+\sigma^2) - \frac{n^2}{2}\mu^2 + (\frac{n^2}{2}-n)\mu^2\right] = \\ &= \sigma^2 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[\bar{s}_{U2}^2] &= \mathbb{E}\left[\left(\frac{1}{\sqrt{n}}(\sum_{i=1}^{\frac{n}{2}} \bar{s}_i - \sum_{i=\frac{n}{2}+1}^{n} \bar{s}_i)\right)\left(\frac{1}{\sqrt{n}}(\sum_{i=1}^{\frac{n}{2}} \bar{s}_i - \sum_{i=\frac{n}{2}+1}^{n} \bar{s}_i)\right)\right] = \\ &= \frac{1}{n}\mathbb{E}\left[\left(\sum_{i\in L}\bar{s}_i - \sum_{i\in R}\bar{s}_i\right)\left(\sum_{i\in L}\bar{s}_i - \sum_{i\in R}\bar{s}_i\right)\right] = \\ &= \frac{1}{n}\mathbb{E}\left[\sum_{i\in L}\sum_{j\in L}\bar{s}_i\bar{s}_j + \sum_{i\in R}\sum_{j\in R}\bar{s}_i\bar{s}_j - \sum_{i\in L}\sum_{j\in R}\bar{s}_i\bar{s}_j - \sum_{i\in R}\sum_{j\in L}\bar{s}_i\bar{s}_j\right] = \\ &= \frac{1}{n}\left[2\left(\frac{n}{2}(\mu^2+\sigma^2) + (\frac{n^2}{4}-\frac{n}{2})\mu^2\right) - 2\left(-\frac{n^2}{4}\mu^2\right)\right] = \\ &= \sigma^2 + n\mu^2 \end{aligned}$$

the rest of the cases can be generalized, where cases k and k+n/2 are the same. then:

$$\begin{aligned}
\mathbb{E}[\bar{s}^2_{Uk}] &= \mathbb{E}\left[\left(\frac{1}{\sqrt{(k-1)(k-2)}}(\sum_{i=1}^{k-2}\bar{s}_i-(k-2)\bar{s}_{k-1})\right)\left(\frac{1}{\sqrt{(k-1)(k-2)}}(\sum_{i=1}^{k-2}\bar{s}_i-(k-2)\bar{s}_{k-1})\right)\right] = \\
&= \frac{1}{(k-1)(k-2)}\mathbb{E}\left[\left((\sum_{i=1}^{k-2}\bar{s}_i-(k-2)\bar{s}_{k-1})\right)\left((\sum_{i=1}^{k-2}\bar{s}_i-(k-2)\bar{s}_{k-1})\right)\right] = \\
&= \frac{1}{(k-1)(k-2)}\mathbb{E}\left[\left(\sum_{i=1}^{k-2}\sum_{j=1}^{k-2}\bar{s}_i\bar{s}_j-2(k-2)\bar{s}_{k-1}\sum_{i=1}^{k-2}\bar{s}_i+(k-2)^2\bar{s}^2_{k-1}\right)\right] = \\
&= \frac{1}{(k-1)(k-2)}\left[(k-2)(\mu^2+\sigma^2)+(k-2)(k-3)\mu^2-(k-2)^2\mu^2+2(k-2)^2(\mu^2+\sigma^2)\right] = \\
&= \frac{1}{k-1}\left[(\mu^2+\sigma^2)+(k-3)\mu^2-2(k-2)\mu^2+(k-2)(\mu^2+\sigma^2)\right] = \\
&= \frac{1}{k-1}\left[\mu^2+(k-3)\mu^2-(k-2)\mu^2+\sigma^2+(k-2)\sigma^2\right] = \sigma^2
\end{aligned}$$

Thus the whole vector is:

$$\mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}}_{\mathbf{U}}] = \begin{bmatrix} \sigma^2 \\ \sigma^2+n\mu^2 \\ \sigma^2 \\ \vdots \\ \sigma^2 \end{bmatrix}$$

A.5 Proof of Proposition 1

$$\begin{aligned}
\mathbb{E}[\mathcal{P}] &= \sigma^2+\frac{1}{(1+F(c)nq)^2}(n\mu^2+\sigma^2)+(n-2)\frac{1}{(1+F(c)n\frac{p+q}{2})^2}\sigma^2 = \\
&= \sigma^2+\frac{1}{(1+u)^2}(n\mu^2+\sigma^2)+(n-2)\frac{1}{(1+w)^2}\sigma^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\mathcal{D}] &= \frac{F(c)nq}{(1+F(c)nq)^2}(n\mu^2+\sigma^2)+(n-2)\frac{F(c)n\frac{p+q}{2}}{(1+F(c)n\frac{p+q}{2})^2}\sigma^2 = \\
&= \frac{u}{(1+u)^2}(n\mu^2+\sigma^2)+(n-2)\frac{w}{(1+w)^2}\sigma^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\mathcal{JC}] &= \frac{(F(c)nq)^2}{(1+F(c)nq)^2}(n\mu^2+\sigma^2)+(n-2)\frac{(F(c)n\frac{p+q}{2})^2}{(1+F(c)n\frac{p+q}{2})^2}\sigma^2 = \\
&= \frac{u^2}{(1+u)^2}(n\mu^2+\sigma^2)+(n-2)\frac{w^2}{(1+w)^2}\sigma^2
\end{aligned}$$

Lets now investigate how each index behaves with change of c (will be done step-by-step to ensure correctness):

First lets compute how $u = F(c)nq$ and $w = F(c)n\frac{p+q}{2}$ behaves in c:

$$\begin{aligned}\frac{\mathrm{d}u}{\mathrm{d}c} =& \frac{\mathrm{d}}{\mathrm{d}c}F(c)nq = \\ =& nq\frac{\mathrm{d}}{\mathrm{d}c}F(c) = \\ =& nqf(c)\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}w}{\mathrm{d}c} =& \frac{\mathrm{d}}{\mathrm{d}c}F(c)n\frac{p+q}{2} = \\ =& n\frac{p+q}{2}\frac{\mathrm{d}}{\mathrm{d}c}F(c) = \\ =& n\frac{p+q}{2}f(c)\end{aligned}$$

Second lets compute how each index behaves in u and w. Since the functional form is the same we can just compute it in general case:

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{(1+x)^2} = -2\frac{1}{(1+x)^3}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{x}{(1+x)^2} = \frac{1-x}{(1+x)^3}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{x^2}{(1+x)^2} = \frac{2x}{(1+x)^3}$$

Now we can put it all together:

Polarization

$$\begin{aligned}\frac{\mathrm{d}\mathbb{E}[\mathcal{P}]}{\mathrm{d}c} &= \frac{\mathrm{d}}{\mathrm{d}c}\left(\sigma^2 + \frac{1}{(1+F(c)nq)^2}(n\mu^2+\sigma^2) + (n-2)\frac{1}{(1+F(c)n\frac{p+q}{2})^2}\sigma^2\right) = \\ &= (n\mu^2+\sigma^2)\frac{\mathrm{d}u}{\mathrm{d}c}\frac{\mathrm{d}}{\mathrm{d}u}\frac{1}{(1+u)^2} + (n-2)\sigma\frac{\mathrm{d}w}{\mathrm{d}c}\frac{\mathrm{d}}{\mathrm{d}w}\frac{1}{(1+w)^2} = \\ &= -2\left(nqf(c)(n\mu^2+\sigma^2)\frac{1}{(1+F(c)nq)^3} + n\frac{p+q}{2}f(c)\sigma^2\frac{1}{(1+F(c)n\frac{p+q}{2})^3}\right) = \\ &= \underbrace{-2nf(c)}_{<0}\left(\underbrace{q(n\mu^2+\sigma^2)}_{>0}\underbrace{\frac{1}{(1+F(c)nq)^3}}_{>0} + \underbrace{\frac{p+q}{2}}_{>0}\sigma^2\underbrace{\frac{1}{(1+F(c)n\frac{p+q}{2})^3}}_{<0}\right) > 0\end{aligned}$$

Thus polarization is decreasing in c

Internal Conflict

$$\begin{aligned}\frac{\mathrm{d}\mathbb{E}[\mathcal{IC}]}{\mathrm{d}c} &= \frac{\mathrm{d}}{\mathrm{d}c}\left(\frac{(F(c)nq)^2}{(1+F(c)nq)^2}(n\mu^2+\sigma^2) + (n-2)\frac{(F(c)n\frac{p+q}{2})^2}{(1+F(c)n\frac{p+q}{2})^2}\sigma^2\right) = \\ &= (n\mu^2+\sigma^2)\frac{\mathrm{d}u}{\mathrm{d}c}\frac{\mathrm{d}}{\mathrm{d}u}\frac{u^2}{(1+u)^2} + (n-2)\sigma\frac{\mathrm{d}w}{\mathrm{d}c}\frac{\mathrm{d}}{\mathrm{d}w}\frac{w^2}{(1+w)^2} = \\ &= \underbrace{2nf(c)}_{>0}\left(\underbrace{q(n\mu^2+\sigma^2)}_{>0}\underbrace{\frac{\overbrace{F(c)nq}^{>0}}{(1+F(c)nq)^3}}_{>0} + \underbrace{\frac{p+q}{2}}_{>0}\sigma^2\underbrace{\frac{\overbrace{F(c)n\frac{p+q}{2}}^{>0}}{(1+F(c)n\frac{p+q}{2})^3}}_{>0}\right) > 0\end{aligned}$$

Thus internal conflict is increasing in c

A.6 Proof of Proposition 3

$$\mathbb{E}[\mathcal{D}] = (n-2)\frac{F(c)n\frac{p}{2}}{(1+F(c)n\frac{p}{2})^2}\sigma^2$$

$$\frac{d\mathbb{E}[\mathcal{D}]}{dc} = \underbrace{n(n-2)f(c)\frac{p}{2}}_{>0}\underbrace{\frac{\overbrace{1-F(c)n\frac{p}{2}}^{?}}{(1+F(c)n\frac{p}{2})^3}}_{>0}\underbrace{\sigma^2}_{>0}$$

, thus:

$$\frac{d\mathbb{E}[\mathcal{D}]}{dc} > 0 \iff 1 - F(c)n\frac{p}{2} > 0 \iff 1 > F(c)n\frac{p}{2} \iff c < Q(\frac{2}{pn})$$

From which it follows that $\mathbb{E}[\mathcal{D}]$ is:

1. increasing in c if $\frac{2}{np} > 1$
2. increasing in $c \in [0, Q(\frac{2}{np}))$, and decreasing in $c \in (Q(\frac{2}{np}), 1]$ if $\frac{2}{np} < 1$

A.7 Proof of Proposition 4

$$\mathbb{E}[\mathcal{P}] = n\mu^2 + 2\sigma^2 + (n-2)\frac{1}{(1+F(c)n\frac{p}{2})^2}\sigma^2$$

$$\mathbb{E}[\mathcal{D}] = (n-2)\frac{F(c)n\frac{p}{2}}{(1+F(c)n\frac{p}{2})^2}\sigma^2$$

$$\mathbb{E}[\mathcal{IC}] = (n-2)\frac{(F(c)n\frac{p}{2})^2}{(1+F(c)n\frac{p}{2})^2}\sigma^2$$

Then

$$\mathbb{E}[\mathcal{W}] = -\left[\alpha(n\mu^2 + 2\sigma^2) + (n-2)\frac{\alpha + \beta F(c)n\frac{p}{2} + (1-\alpha-\beta)(F(c)n\frac{p}{2})^2}{(1+F(c)n\frac{p}{2})^2}\sigma^2\right]$$

lets denote $F(c)n\frac{p}{2}$ by u, then:

$$\mathbb{E}[\mathcal{W}] = \alpha(n\mu^2 + 2\sigma^2) + (n-2)\frac{\alpha + \beta u + (1-\alpha-\beta)(u)^2}{(1+u)^2}\sigma^2$$

Note that:

$$\frac{d\mathbb{E}[\mathcal{W}]}{dc} = \frac{du}{dc}\frac{d\mathbb{E}[\mathcal{W}]}{du}$$

$$\frac{du}{dc} = f(c)n\frac{p}{2}$$

$$\frac{d\mathbb{E}[\mathcal{W}]}{du} = -\frac{u(-2\alpha - 3\beta + 2) - 2\alpha + \beta}{(1+u)^3}$$

Then:

$$\frac{d\mathbb{E}[\mathcal{W}]}{dc} = -f(c)n\frac{p}{2}\frac{F(c)n\frac{p}{2}(-2\alpha - 3\beta + 2) - 2\alpha + \beta}{(1+F(c)n\frac{p}{2})^3}$$

First lets find the critical points for $c \in [0,1]$:

$$\frac{d\mathbb{E}[\mathcal{W}]}{dc} = 0 \iff -f(c)n\frac{p}{2}\frac{F(c)n^{\frac{p}{2}}(-2\alpha-3\beta+2)-2\alpha+\beta}{(1+F(c)n^{\frac{p}{2}})^3} = 0 \iff$$
$$F(c)n\frac{p}{2}(-2\alpha-3\beta+2)-2\alpha+\beta = 0 \iff c^* = Q\left(\frac{2\alpha-\beta}{n^{\frac{p}{2}}(2-2\alpha-3\beta)}\right)$$

Now we need to ensure concavity in c^* if it is a maximum

$$\frac{d^2\mathbb{E}[\mathcal{W}]}{dc^2} = \frac{d}{dc} - f(c)n\frac{p}{2}\frac{F(c)n^{\frac{p}{2}}(-2\alpha-3\beta+2)-2\alpha+\beta}{(1+F(c)n^{\frac{p}{2}})^3} =$$
$$-\left(f'(c)n\frac{p}{2}\frac{F(c)n^{\frac{p}{2}}(-2\alpha-3\beta+2)-2\alpha+\beta}{(1+F(c)n^{\frac{p}{2}})^3} + (f(c)n\frac{p}{2})^2\frac{F(c)n^{\frac{p}{2}}(4\alpha+6\beta-4)+4\alpha-6\beta+2}{(1+F(c)n^{\frac{p}{2}})^4}\right)$$

Now by plugging the c^* in relevant places we get:

$$\frac{d^2\mathbb{E}[\mathcal{W}(c^*)]}{dc^2} = -f'(c^*)n\frac{p}{2}\frac{\frac{2\alpha-\beta}{n^{\frac{p}{2}}(2-2\alpha-3\beta)}n^{\frac{p}{2}}(-2\alpha-3\beta+2)-2\alpha+\beta}{(1+F(c^*)n^{\frac{p}{2}})^3} -$$
$$(f(c^*)n\frac{p}{2})^2\frac{\frac{2\alpha-\beta}{n^{\frac{p}{2}}(2-2\alpha-3\beta)}n^{\frac{p}{2}}(4\alpha+6\beta-4)+4\alpha-6\beta+2}{(1+F(c^*)n^{\frac{p}{2}})^4} =$$
$$-(f(c^*)n\frac{p}{2})^2\frac{-4\beta+2}{(1+F(c^*)n^{\frac{p}{2}})^4}$$

Thus it is a maximum for $\beta < \frac{1}{2}$ as the function is continuous and concave in $c^* \in [0,1]$ there.

Then we need to check where $\frac{2\alpha-\beta}{n^{\frac{p}{2}}(2-2\alpha-3\beta)}$ is in $[0,1]$ as it is an argument of the quantile function, while keeping $\beta < \frac{1}{2}$.

Note that $\alpha < 1-\beta$ by definition

$$\begin{cases} \frac{2\alpha-\beta}{n^{\frac{p}{2}}(2-2\alpha-3\beta)} > 0 \\ \alpha < 1-\beta \\ \beta < \frac{1}{2} \end{cases} \iff \begin{cases} 2\alpha-\beta > 0 \\ 2-2\alpha-3\beta > 0 \\ \alpha < 1-\beta \\ \beta < \frac{1}{2} \end{cases} \text{ or } \begin{cases} 2\alpha-\beta < 0 \\ 2-2\alpha-3\beta < 0 \\ \alpha < 1-\beta \\ \beta < \frac{1}{2} \end{cases} \iff$$

$$\begin{cases} \alpha > \frac{\beta}{2} \\ \alpha < 1-\frac{3}{2}\beta \\ \beta < \frac{1}{2} \end{cases}$$

As the second case is a contradiction due to $2\alpha-\beta>0$ and $\beta<\frac{1}{2}$ being not possible at the same time

Now keeping in mind the conditions of $\frac{2\alpha-\beta}{n\frac{p}{2}(2-2\alpha-3\beta)}$ being positive, lets check when it is below 1:

$$\frac{2\alpha-\beta}{n\frac{p}{2}(2-2\alpha-3\beta)}<1 \iff 2\alpha-\beta<n\frac{p}{2}(2-2\alpha-3\beta) \iff 2\alpha+n\frac{p}{2}2\alpha<n\frac{p}{2}(2-3\beta)+\beta \iff$$
$$2\alpha(1+n\frac{p}{2})<n\frac{p}{2}(2-3\beta)+\beta \iff \alpha<\frac{n\frac{p}{2}(2-3\beta)+\beta}{2(1+n\frac{p}{2})}=\frac{\beta(1-3n\frac{p}{2})+np}{2(1+n\frac{p}{2})}$$

Where the RHS for $\beta,\frac{1}{2}$ is always below (and converges in $n\frac{p}{2}$ to) $1-\frac{3}{2}\beta$, thus the optimal c^* is not a corner solution for:

$$\begin{cases}\alpha>\frac{\beta}{2}\\ \alpha<\frac{n\frac{p}{2}(2-3\beta)+\beta}{2(1+n\frac{p}{2})}\\ \beta<\frac{1}{2}\end{cases}$$

here it is important to note that $\beta<\frac{1}{2}$ is redundant as the two above conditions ensure it Now we can check when the optimal censor point is either 0 or 1. Lets start by computing the values of $\mathbb{E}[\mathcal{W}]$ at $c=1$ and at $c=0$:

$$\mathbb{E}[\mathcal{W}(0)]=-\alpha(n\mu^2+2\sigma^2)-\alpha(n-2)\sigma^2$$
$$\mathbb{E}[\mathcal{W}(1)]=-\alpha(n\mu^2+2\sigma^2)-(n-2)\frac{\alpha+\beta n\frac{p}{2}+(1-\alpha-\beta)(n\frac{p}{2})^2}{(1+n\frac{p}{2})^2}\sigma^2$$

Then:

$$\mathbb{E}[\mathcal{W}(0)]>\mathbb{E}[\mathcal{W}(1)] \iff$$
$$-\alpha(n\mu^2+2\sigma^2)-\alpha(n-2)\sigma^2>-\alpha(n\mu^2+2\sigma^2)-(n-2)\frac{\alpha+\beta n\frac{p}{2}+(1-\alpha-\beta)(n\frac{p}{2})^2}{(1+n\frac{p}{2})^2}\sigma^2 \iff$$
$$\alpha(n-2)\sigma^2<(n-2)\frac{\alpha+\beta n\frac{p}{2}+(1-\alpha-\beta)(n\frac{p}{2})^2}{(1+n\frac{p}{2})^2}\sigma^2 \iff \alpha<\frac{\alpha+\beta n\frac{p}{2}+(1-\alpha-\beta)(n\frac{p}{2})^2}{(1+n\frac{p}{2})^2} \iff$$
$$\alpha+2\alpha n\frac{p}{2}+\alpha(n\frac{p}{2})^2<\alpha+\beta n\frac{p}{2}+(1-\alpha-\beta)(n\frac{p}{2})^2 \iff 2\alpha(1+n\frac{p}{2})<\beta(1-n\frac{p}{2})+n\frac{p}{2} \iff$$
$$\alpha<\frac{\beta(1-n\frac{p}{2})+n\frac{p}{2}}{2(1+n\frac{p}{2})}$$

Thus summing up:

In the highly homophilic case the c that maximizes the Welfare is:

1. $c^* = 1$, if $\alpha > \frac{\beta(1-3n\frac{p}{2})+np}{2(1+n\frac{p}{2})}$ and $\alpha > \frac{\beta(1-n\frac{p}{2})+n\frac{p}{2}}{2(1+n\frac{p}{2})}$
2. $c^* = Q(\frac{2}{np}\frac{2\alpha-\beta}{2-2\alpha-3\beta})$, if $\alpha > \frac{\beta}{2}$ and $\alpha < \frac{\beta(1-3n\frac{p}{2})+np}{2(1+n\frac{p}{2})}$
3. $c^* = 0$, if $\alpha < \frac{\beta}{2}$ and $\alpha < \frac{\beta(1-n\frac{p}{2})+n\frac{p}{2}}{2(1+n\frac{p}{2})}$