Understanding Social Media: Misinformation, Attention, and Digital Advertising

by

James Siderius

A.B., Princeton University (2016)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2023

Author . . . .
Department of Electrical Engineering and Computer Science
September 19, 2022

Certified by. . . . .
Asuman Ozdaglar
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Understanding Social Media: Misinformation, Attention, and Digital Advertising

by

James Siderius

A.B., Princeton University (2016)

S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Electrical Engineering and Computer Science
on September 19, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Online platforms have fundamentally changed the dynamics of social interactions and information transmission. In this thesis, I explore recent trends in social media through models and experiments of user behavior, platform algorithms and incentives, and policy initiatives. I focus on the social consequences of new communication technologies, their intended and unintended societal consequences, and how to steer them in more socially beneficial directions.

In recent years, social media has become a breeding ground for misinformation, but the reasons misinformation spreads are still imperfectly understood. First, I discuss the role of social media in the propagation of misinformation, how latent platform algorithms may exacerbate its influence, and analyze various policies to correct misinformation spread. Technological advances stemming from social media have also enabled users to systematically access a deluge of information; yet, it is unclear to what extent this technology has actually helped to better inform. In the second part of the thesis, to characterize the landscape of digital content, I propose a model of content creation and consumption on digital platforms where users have limited attention, and discuss related experiments on the role of algorithmic ranking in user engagement. Lastly, we observe that business models of online platforms drive much of the content creation and algorithmic choices of platforms, and ultimately impact human-machine interactions. The final part of the thesis discusses the various business models of media platforms, their implications for consumer welfare, and possible remedies.

Thesis Supervisor: Asuman Ozdaglar
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank all of my family, friends, colleagues, and mentors for their support and guidance throughout my PhD tenure.

First, I would like to give my deepest thank you to my wife, Jackie Dewart, for her unwavering support even in the most difficult times throughout the last six years. I am also greatly appreciative of my parents, Cheryl and Martin Siderius, and my sister, Kristen Siderius, who have always encouraged me to pursue my academic passions.

I'd also like to thank my close personal friends, Nour Alharithi, Yem Alharithi, Daway Chou-Ren, Darshan Desai, Alex Dominguez, Dan Jang, Aaron Katz, Davy Perlman, Paul Phillips, Alex Wollack, and Eddie Zhou, among others, for their constant encouragement and the much-needed distractions they offered over the last six years.

I am also grateful for the close colleagues and friends I had at MIT and in LIDS, specifically Jason Altschuler and Nuri Denizcan Vanli, as well as various co-authors, Sarah Cen, Nicole Immorlica, Brendan Lucier, Markus Mobius, Mohamed Mostagir, and Alireza Tahbaz-Salehi, for their collaboration on fascinating topics, and who have helped me learn deeply. I'd also like to thank Charles Lyu, who has been a great help in pushing many of our experimental projects.

Finally, I would like to thank my thesis readers and advisors, Daron Acemoglu, Adam Berinsky, Daniel Huttenlocher, and Asuman Ozdaglar. Adam and Dan have been very influential in broadening the potential vision and impact of our research, both experimentally and theoretically. Likewise, and quite importantly, I cannot express enough gratitude to my advisors, Daron Acemoglu and Asuman Ozdaglar, who have been a strong and guiding force throughout the work of this thesis (among other projects). It has been a great pleasure to learn from both of them over the years. I can comfortably say that this thesis would not have been possible without their meaningful mentorship.

I would also like to acknowledge the financial support for my research, graciously provided by the Army Research Office grant on "Toward Large-Scale Human Interactions."

Contents

Chapter 1

Introduction

This chapter introduces the main ideas of the thesis as well as provides some essential background. While the focus of this thesis is on social media, many of the ideas apply more broadly to online media platforms (especially Chapter 5). In Section 1.1, I briefly describe the nature of various online media platforms, including both social media and other emerging media platforms. A main focus is on models and experiments of social media related to the spread of misinformation (discussed in both Chapters 2 and 3). In Section 1.2, I define what misinformation is, provide a brief history that predates social media, and give an overview of how social media has changed the paradigm for why and how misinformation spreads. Lastly, in Section 1.3, I informally discuss the role of platform incentives and algorithmic choices on societal outcomes (which are studied throughout in greater detail), and give a short overview of regulatory solutions that have been discussed recently in the public sphere.

1.1 Online Media Platforms

Online media has transformed many aspects of modern life. The overwhelming majority of Americans now rely on online platforms for media of all kinds, including traditional entertainment, such as TV and movies,[1] as well as social interaction, which has become increasingly more virtual with the rise of social media sites such as Facebook and Twitter.[2] The foreseen benefits

[1] Much evidence shows consumers have substituted away from classic cable service in favor of streaming services such as Netflix, Hulu, and HBO (among many others), see Abreu et al. (2017) and Arditi (2021).

[2] It has been well-documented (e.g., in Goel and Gupta (2020) and Cinelli et al. (2020)) that the COVID-19 pandemic has also accelerated much of the transition to online social interaction.

of these platforms are clear. First, they provide less friction in matching content to individual users and their preferences. Users can quickly and easily access content that might interest them, which allows for a broader selection of sources tailored to a range of idiosyncratic interests. Second, there are fewer hurdles to releasing and disseminating content; this includes potentially fewer barriers to entry for content providers and offers a more transparent channel for free speech and democratic discourse. Lastly, it also allows platforms to engineer rich algorithms intended to further facilitate user experiences and optimize personal interactions. The focus of this thesis is to explore the hidden consequences of these new technologies, whether these innovations deliver on their perceived benefits, and how they might introduce negative societal repercussions (either intended or unintended).

1.1.1 Social Media

Figure 1-1. Some of the most popular social media sites (Twitter, Reddit, Facebook), as well as fringe social media focused around extremist ideas (Parler, Trump Social).

Social media has played an increasingly important role globally in the last several decades, acting as an integral part of the lives of billions of people worldwide. There are two types of actors on social media: users and the platform. Users benefit from social media via information and entertainment. They may learn more about a particular political issue, discover new vacation spots from a friend who posted photos in Belize, watch an amusing cat video, or come across a new product being actively advertised. At the core of social media sites are recommendation algorithms that determine the likelihood of individuals seeing different types of information, which influences user interaction. Although these algorithms are designed by platforms with specific objectives, researchers and policymakers have recently started recognizing that some of their consequences may be socially problematic, including echo chambers, misinformation spread, and negative effects on the mental health of vulnerable populations. These pernicious outcomes may be amplified further by the business models of social media companies who are monetized by digital advertising and incentivized to promote digital addiction (see Allcott et al. (2020) and Allcott et al. (2022)).

In the early 2000s, social media sites like MySpace and Facebook were founded on the basis of facilitating social interactions and boosting connectivity from all parts of the world.[3] Soon after, social media sites also evolved to serve other purposes, such as the quick and abundant dissemination of news and entertaining content (see Cuthbertson et al. (2015) and Edosomwan et al. (2011)). In many ways, social media paved the way for easier communication, more democratic discourse, and an exposure to broader and more diverse viewpoints (see, for example, Wilson and Stock (2021)).

However, under the surface, the effects of social media are much more subtle. Greater access to information and the ability to more easily form social connections have also lead to the formation of one-sided communities that act as echo chambers with little diversity of opinion (as evidenced by Quattrociocchi et al. (2016), Mosleh et al. (2021b), and Bakshy et al. (2015)). In fact, self-segregation on social media is one of the leading theories for accentuated polarization and division in this United States in the last twenty years (according to Sims and Grant (2021) and Pew Research Center (2014)). Moreover, there is substantial evidence that social media can have negative effects on the formation of social relationships and has caused a decline in mental health, especially among adolescents (e.g., see Bashir and Bhat (2017), Berryman et al. (2018), and Strickland (2014)). Similarly, on a platform where content can be generated and posted at almost no cost, social media may act as a medium for the propagation of "news" that contains little to no information, misleading ideas, or outright false claims ("misinformation"). It has been suggested that viral misinformation can even influence the actions of those who consume it, potentially to the point of affecting democratic decisions (see Allcott and Gentzkow (2017)).

These negative societal consequences are in part due to the nature of social media compared to traditional media. Traditional media typically consists of a handful of news outlets with established reputability, who publish content for mass consumption.[4] These outlets are incentivized to target different, but coarse sectors of the consumer market (who might have heterogenous preferences) and maintain a reputation for quality and informativeness. Social

[3] According to McWilliams (2009), MySpace's slogan was "a place for friends" and Facebook's was "to help connect and share with people in your life".

[4] For models of content production (and bias) in traditional media settings, see Gentzkow and Shapiro (2006) and Allon et al. (2021).

media fundamentally differs on two dimensions. First, content can be more easily posted and shared on social media, often with little repercussions for spreading false content or misinformation. This is largely due to a lack of reputational concerns when (mis)information can be effortlessly created and disseminated under various aliases.[5] Secondly, the abundance of content on social media allows users to easily self-select into very niche communities that in general *decrease* exposure to broader mainstream ideas.

The business model of social media companies does not (by any means) help mitigate these negative consequences, and may even accentuate them. Social media platforms generate most of their revenue through digital advertising, which is more valuable the longer users stay on the platform. Recent empirical work has shown that this can create a perverse incentive for platforms to get social media users "hooked" on the site (see Allcott et al. (2020) and Allcott et al. (2022)). Similarly, it has been shown that social media recommendation algorithms can increase user engagement specifically by pushing sensational content (often likely to contain misinformation) within ideologically-congruent populations (e.g., as in Vosoughi et al. (2018) and Levy (2021)). At the same time, the overload of content on social media pushes the platform toward suggesting catchy content, perhaps with little to no informational value, because this is most effective for capturing attention. On top of this, because digital advertising is the main business model for social media platforms, targeting susceptible populations with manipulative ads can be the most profitable way to attract advertisers and increase revenue.

1.1.2 Other Online Media: WhatsApp, Yelp, YouTube, and TikTok

Figure 1-2. Other media platforms which have recently become popular.

In recent years, online media platforms besides social media have also expanded and become quite popular. WhatsApp, a common platform used in countries outside of the US,

[5] For example, a Macedonian teenager spread misinformation during the run-up to the 2016 U.S. presidential election by designing a website that mimicked the Huffington Post (see Allcott and Gentzkow (2017)), and faced no fallout from the blatant fabrication of content masquerading as news.

has allowed people to connect more easily where cell towers are less available. Yelp, among other sites such as Amazon and Airbnb, have developed platforms for crowdsourced reviews of restaurants, products, and vacation spots, to name just a few. YouTube has become a dominant player in streaming online entertainment, along with other services such as Netflix and Hulu. Most interesting, arguably, are apps such as Snapchat and TikTok, with limitations on how content can be shared (e.g., 10-second photos for Snapchat or 30-second videos for TikTok) which have taken off explosively especially among younger demographics.

As mentioned before, while the vast majority of social media sites are monetized fully through digital advertising, other media platforms have more nuanced business models. WhapsApp and YouTube both offer freemium services, where a user can pay a subscription fee to avoid advertisements but can otherwise enjoy the site for free (with ads). Other platforms such as Yelp (and Google) solicit bids from clients to provide preferential ranking in searches from its users. Business models of platforms like Amazon and Airbnb, however, rely largely on selling the products or vacation spots advertised on the site.

Many of the societal concerns about social media extend to this broader setting as well. WhatsApp had been used a way to spread lies about COVID-19 and other misinformation in India, where there is less access to formal education and fact-checking resources (Neyazi et al. (2021), Garimella and Eckles (2020), and Banaji et al. (2019)). Similarly there is both strong empirical and theoretical evidence that the fake reviews on Yelp can be extensive, and hurt both consumer welfare and the welfare of restaurants who do not participate in the fraudulent activity (Luca and Zervas (2016) and Mostagir and Siderius (2022c)). As mentioned before, the business models of social media and more recent platforms like TikTok or Instagram may even encourage the platform to create insecurity and depression among its users (Catlett (2022) and Giordano et al. (2022)).

1.1.3 Thesis Overwiew and Organization

A main goal of this thesis is to develop workable approaches to understand the interplay of large numbers of stakeholders with different beliefs and interests, which calls for a dynamic game-theoretic approach. Though challenging, this step is critical for deriving general insights and new empirical predictions. I develop both new insights and new predictions about the consequences of content creation, platform recommendation algorithms, and business models developed and operated by social media platforms interested in maximizing engagement. With these predictions in hand, I propose ongoing work to empirically test user behavior under different algorithms to better calibrate the underlying parameters of our models and identify the key drivers of online behavior. Leveraging our existing modeling framework that reveals the tension between users and the platform (and society at large), I also study the consequences of different policy interventions, including content moderation efforts, information tagging (e.g., the provenance of different messages), and algorithmic network regulation. Ongoing work (outside this thesis) hopes to extend the analysis of these policies to metrics other than misinformation, such as algorithmic fairness in the context of social media.

The basis of many of the predictions is a conceptual framework in which users with different ideological leanings consume information and communicate with each other. Via their choice of recommendation algorithm, a social media platform determines the likelihood of communication between any two pairs of agents (and the likelihood that they will see certain news items). The platform's objective is to maximize user engagement, which then enables more profitable digital ad targeting. The important lesson from this framework is a specific set of predictions on when platforms will propagate low-reliability content and especially do so by creating filter bubbles (artificial echo chambers). Such filter bubbles increase the spread of questionable content among like-minded users, who then end up engaging more with this content, creating profitable ad targeting opportunities. The resulting pattern of communication and belief formation may be at odds with various social objectives, especially because misinformation spreads quite strongly. What can be done in order to align platform algorithms with social objectives? Some policies look appealing at first, but their effects turn out to be more complex. For example, tagging a subset of misinformation so that users are aware when they are engaging with questionable content could be useful, but under certain circumstances it can also backfire because of an "implied truth effect": knowing that some misinformation has been tagged, users are now more optimistic about the quality of the information they receive, even if it is unreliable (see Pennycook et al. (2018)). One of our objectives is to experimentally investigate these issues and thus generate insights on better policy design.

Another major focus of this thesis is in understanding the algorithms on social media and their impact on user behavior. Toward this goal, I also investigate this issue experimentally by designing a new interface where the implications of different recommendation algorithms can be traced. Participants will be given a feed of news items that they can like or share. These feeds are sorted by simple ranking algorithms prioritizing articles that cater to user preferences, articles shared by friends, or both. We will then measure the impact of different ranking methods on user behavior, such as likes, shares and dwell time, compared to the baseline of a randomized content feed. This allows us to understand platform incentives for ranking, observe whether ranking algorithms tend to push more exploitative content, and gain insights into the role that algorithms play in undesirable social consequences. We also strive to understand the role that user attention plays in not just platform recommendations, but also content creation. This offers a novel analysis of competition for user attention. In particular, in environments with limited attention, different content will have to compete not just for the monetary resources of users, but even more importantly for their attention. This type of competition can be a conduit to a race to the bottom. Catchy content that is mostly click-bait may become more attractive under information overload. This prediction will be investigated experimentally in our new testbed.

In this thesis, I explore whether the benefits of social media have indeed materialized, explore the more nuanced impacts of digital media platforms, and consider various regulatory solutions to correct any potential harms introduced. At the core of these considerations are the following questions:

1. Does more information mean better information?
2. How do the business models of online platforms shape algorithms and create incentives to show certain types of content?
3. Can we more closely align platform and user behavior with societal objectives through regulation of online interactions?

The thesis is organized as follows. Chapters 2 and 3 discuss misinformation on online platforms and their broader impact on society. Chapter 2 focuses on the useful benchmark of strategic (Bayesian) agents in a social media environment with a platform who wants to maximize user engagement (and revenue). Chapter 3 considers alternative models that relax the strategic nature of social media agents and emphasizes potential behavioral reasons for the spread of misinformation. Chapter 4 then endogenizes content creation to understand incentives for providing content when there is information overload and competition for user attention. Finally, Chapter 5 shifts the focus to business models of online platforms, specifically digital advertising, and analyzes the welfare implications for platform users.

1.2 Misinformation in the Digital Era

Misinformation has had a profound impact on recent society. What exactly is misinformation? In this thesis, we adopt a broad definition that captures many nuances, but with the same basic concept. Misinformation is content with the potential to deceive. This can, of course, include *disinformation*, which is purposefully fabricated content with the intention of misleading. The story of the Macedonian teenager masquerading as the Huffington Post is one such example (see Footnote 5), and propaganda spread by a government is another.

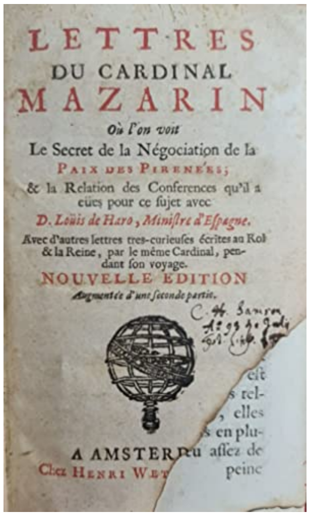
There are other forms of misinformation (not intentionally created and spread), typically dubbed "fake news." For example, conspiracy theories (e.g., Pizzagate, QAnon, Flat Earth) may not be spread to deliberately deceive, yet they contain "alternative facts" known to be false. There are less blatant fake news articles that might even appear on the surface to be true, but are often called out by independent fact-checkers such as Snopes.[6] These would all fall under the umbrella as misinformation. However, we also define misinformation to include all types of misleading content, even if the facts are *correct* but presented in a *misleading* way. For instance, a headline that says "3-year old, youngest person to die with COVID-19" may be technically accurate (indeed the infant had COVID-19), but the child also had a heart condition largely responsible for their death, and so this would be classified as misinformation under my definition.

1.2.1 A Brief History of Misinformation

Misinformation is not a new phenomenon. Throughout history, new communication technologies have come with both positive and negative consequences. On the positive end, they have helped more easily disseminate profound ideas and inform the broader population. However, the flip side of the coin is that new technology has simultaneously supplied heretics with a

[6] At the time of writing this, "Starbucks is Going Cashless in UK, US, and Canada?" was the top false headline on Snopes, see `https://www.snopes.com/fact-check/starbucks-cashless/`.

means for spreading false ideas or “misinformation.” For example, the printing press was a major development that led to widespread literacy throughout Europe and in many ways kickstarted the Renaissance. At the same time, it was also used to pass pamphlets during the French Revolution with blatantly false rumors about Louis XIV’s Cardinal Mazarin’s treason, often advocating for his execution (see Figure 1-3).



LETTRES
DU CARDINAL
MAZARIN
Où l'on voit
Le Secret de la Négociation de la
PAIX DES PIRENÉES;
& la Relation des Conferences qu'il a
eües pour ce ſujet avec
D. Loüis de Haro, Miniſtre d'Eſpagne.
Avec d'autres lettres tres-curieuſes écrites au Roi
& la Reine, par le même Cardinal, pen-
dant ſon voyage.
NOUVELLE EDITION
Augmentée d'une ſeconde partie.

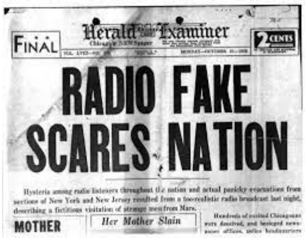
C. H. Sanron [illegible]

eſt
s tel-
, elles
s en plu-
u aſſez de
peine

A AMSTER
Chez HENRI WET

Figure 1-3. Pamphlet circulating during the French Revolution, circa 1693.

The invention of the radio also had undeniable widespread implications for communication, such as the smooth orchestration of orders during the first and second world wars. However, during the same era, it was also used widely for communicating propaganda and spreading xenophobic ideas meant to incite hate and bigotry. In one of the most famous examples, college students used a radio broadcast in 1938 in an attempt to incite fear about gas explosions that occurred on Mars (see Figure 1-4).



FINAL

Herald Examiner

Chicago's NEWSpaper

2 CENTS

MONDAY—OCTOBER 31—1938

RADIO FAKE SCARES NATION

Hysteria among radio listeners throughout the nation and actual panicky evacuations from sections of New York and New Jersey resulted from a too-realistic radio broadcast last night, describing a fictitious visitation of strange men from Mars.

MOTHER

Her Mother Slain

Hundreds of excited Chicagoans were deceived, and besieged newspaper offices, police headquarters

Figure 1-4. Newspaper reporting misinformation spread over radio, circa 1938.

Our society now faces a new technological advance in communication, social media, and it is not obvious what to what extend the positive and negative impacts of this technology will be. In many ways, it is quite similar to previous innovations, and similar limitations on free speech (e.g., libel) need to be imposed on a grander scale to prevent the adoption of false ideas. However, in other ways, we are entering uncharted waters, with social media fundamentally impacting the entire landscape of content, including how misinformation spreads and how their fundamental business model impacts consumer welfare.

1.2.2 Misinformation on Social Media

Social media has become a major source of information for many Americans. Leading up to the 2016 US presidential election, around 14% of Americans indicated social media as their primary source of news (Allcott and Gentzkow (2017)), and by 2019, over 70% of Americans reported receiving at least *some* of their news from social media (Levy (2021)). At the same time, there is growing concern about misinformation in social media, including made-up news stories such as those claiming that there were no mass shootings under Donald Trump or that Hillary Clinton approved ISIS weapon sales.[7] Some recent evidence also suggests that misinformation on social media has impacted critical decisions such as vaccinations against COVID-19 (see Pennycook et al. (2018); Pennycook et al. (2020b)).

Although there is yet no consensus on what promotes the spread of falsehoods and misleading content on social media, two sets of factors have been emphasized. The first is the presence of echo chambers, which arise when individuals communicate and share content with like-minded users (Sunstein (2018), Lazer et al. (2018)). Törnberg (2018) and Vicario et al. (2016) show that echo chambers reinforce existing political viewpoints and tend to propagate misinformation. Social media, much more than traditional media, allows individual users to choose who and what they listen to, and thus echo chambers may be an unavoidable side effect. However, there is also evidence that echo chambers are a result of the "filter bubbles" that platform algorithms create (see Levy (2021) on Facebook). The second factor conjectured to have fueled misinformation is the general political polarization in many countries, and especially the United States,[8] and there is some preliminary evidence suggesting that polarization has indeed contributed to selective exposure to questionable content on social media as well (Guess et al. (2018)). Despite the importance and salience of these issues, we do not currently have a

[7] See `https://www.snopes.com/fact-check/mass-shootings-under-trump/` and `https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html`, respectively.

[8] While there has been some debate about whether polarization has been mainly among politicians (see Fiorina et al. (2008) and Prior (2013)), there is considerable evidence that polarization has also risen among the general public (see Pew Research Center (2014) and Abramowitz (2010)).

framework to understand how online interactions impact the spread of misinformation and what factors shape incentives for sharing low-reliability content.

Unfortunately, misinformation has become an inextricable component of how people learn about the world and make decisions. Persistent disagreement over objective facts is becoming increasingly commonplace (Alesina et al., 2020; Bursztyn et al., 2020), leading to the assertion that we currently live in a post-truth, "alternative facts" world (McIntyre, 2018). This makes it particularly difficult when society faces a collective action problem whose outcome depends on a substantial proportion of the population agreeing to take a specific action, e.g., vaccination or wearing a mask. Indeed, Loomba et al. (2021) show that in the U.S. and U.K., an average decline of 6.2 percentage points in the acceptance of the COVID-19 vaccine is directly attributable to misinformation.

What makes people believe false claims? A growing empirical and experimental literature argues that it depends on their cognitive sophistication.[9] Pennycook and Rand (2019) show that more sophisticated agents are less likely to fall for misinformation. At the same time, there is ample evidence that sophisticated agents are more likely to disagree over objective facts.[10] This presents an interesting puzzle: if sophisticated agents are more likely to learn the truth, then why do we observe more disagreement within that group over what the truth is? Competing explanations argue that this disagreement arises from political polarization and partisan bias (e.g., Taber and Lodge (2006); Taber et al. (2009); Kahan et al. (2017)), or can simply be explained as an outcome of unbiased Bayesian reasoning (e.g., the recent experiments in Tappin et al. (2020)).

Throughout this thesis, I will provide a framework to better understand these deep questions, through both models and experiments on social media.

[9] From Tappin et al. (2020): "indicators of cognitive sophistication [can include] educational attainment, science literacy, numeracy, specific topic knowledge, and a propensity for analytic thinking." This is also correlated with performance on the Cognitive Reflection Test (Pennycook and Rand, 2019). Pennycook et al. (2021b) show that getting people to think more carefully about the accuracy of the news they read can lead to better discernment of false information.

[10] This relationship between increased cognitive sophistication and disagreement is widely documented across multiple issues; see for example Drummond and Fischhoff (2017); Kahan et al. (2012); Hamilton et al. (2015).

1.3 Social Media AI and Regulation

At the heart of many of these problems are social media algorithms often driven by financial objectives. While discussed more in-depth in later chapters, we provide a short summary of the main platform incentives that drive the key findings of the thesis. We then briefly survey some of the more recent policy suggestions to better align these incentives with societal objectives.

1.3.1 Platform Incentives and Algorithms

A key difference between traditional social interactions (or traditional media consumption) and social media interactions (or online media consumption) is the existence of a platform, which has distinct incentives from its users (or society). Platforms often employ latent algorithms that better align online interactions with their own objectives, often at the cost of the users' experience and society at large. In this thesis, we focus on three separate but related platform incentives:

(i) In Chapters 2 and 3, we consider a social media environment where the platform's main objective is to maximize user engagement with the platform, but is indifferent between whether content contains misinformation or is truthful. This objective is a proxy for how most social media companies make revenue, through advertising, which is more profitable the longer users stay on the platform. There is an obvious disconnect between platform incentives and societal objectives: the platform is likely to design algorithms that prevent misinformation from being detected and has a higher likelihood to go "viral". We study this phenomenon specifically in Section 2.1.

(ii) In Chapter 4, we consider the role of "news feeds" with platform-recommended content. When users have limited attention and there is competition for content viewership, the platform may be incentivized to prioritize catchy and stimulating content over informative articles. This implies that while the platform provides the technological capacity for more information to appear on the platform, content creators and platforms alike are encouraged can often tend to push lower quality content.

(iii) In Chapter 5, we turn our attention to one of the central business models of social media, digital advertising. In particular, we look at the negative welfare implications that might arise when platforms specifically target ads at susceptible populations and leverage different business models (such as freemium) to maximize revenue at the cost of consumer manipulation.

1.3.2 Algorithmic Solutions

In 1949, the FCC introduced a policy known as the "fairness doctrine" which required news providers to present both sides of a controversial issue (see Simmons (1976)). This policy was eliminated in 1987, giving rise to news outlets that were heavily one-sided and paving the way for partisan talk radio (Clogston (2016)). While one-sided news does not necessarily contain misinformation, it provides an avenue for presenting content that is skewed toward one perspective, with misinformation (when it exists) also likely arising from this perspective. For example, a liberal media outlet that is not required to present diverse content is more likely to present misinformation in favor of left-wing political candidates. Hence, we assume a policy that requires "equal coverage" of both sides of a topical issue, and model this as content being less likely to present strongly misleading information toward one perspective.

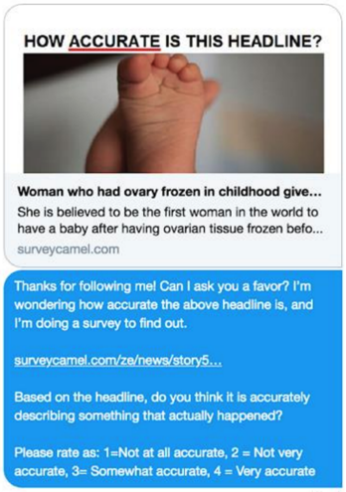
A modern-day equivalent of the fairness doctrine is the idea of requiring social media platforms to provide more diverse news feeds. As seen in Levy (2021), platforms typically try to increase user engagement by (algorithmically) recommending stories that match users' profiles, and this can result in "filter bubbles" that limit the scope of counter-attitudinal content that users see. This has been linked to the propagation of misinformation and its influence on outcomes such as the 2016 presidential race (see Allcott and Gentzkow (2017)). A proposed solution is to regulate these algorithms in order to provide more diverse news, for example by requiring content be shown "uniformly at random" from one's social network, and not selectively filtered (e.g., Sunstein (2018); Cen and Shah (2020)). This is similar to an equal-coverage policy, where users of the social media platform ideally learn from a variety of sources with different perspectives.

1.3.3 Content Moderation and Censorship

Censorship is one of the oldest policies for controlling access to information. The practice is controversial because it typically involves a unilateral decision (e.g., by a platform or a government) to remove certain information and make it inaccessible. The use of the policy in modern times is more complicated as platforms try to regulate content by balancing freedom of expression with reducing harmful speech, e.g., the aforementioned example of Twitter removing COVID-19 misinformation from the platform.

Other more mild forms of content moderation have been implemented on social media sites such as Reddit and Twitter. On Reddit, quarantines are often placed on communities to simply make joining them more difficult and to prevent posts from the page reaching other more mainstream ones. Similarly, Twitter has implemented policies that rank tweets or hashtags affiliated with misinformation lower even if the search terms are a strong fit.[11] Regulatory solutions of this nature can often strike a nice middle ground between a fully open platform and one where the platform can simply remove content at will via censorship.

1.3.4 Nudging and Provenance



HOW ACCURATE IS THIS HEADLINE?

Woman who had ovary frozen in childhood give...

She is believed to be the first woman in the world to have a baby after having ovarian tissue frozen befo...

surveycamel.com

Thanks for following me! Can I ask you a favor? I'm wondering how accurate the above headline is, and I'm doing a survey to find out.

surveycamel.com/ze/news/story5...

Based on the headline, do you think it is accurately describing something that actually happened?

Please rate as: 1=Not at all accurate, 2 = Not very accurate, 3= Somewhat accurate, 4 = Very accurate

Figure 1-5. A sample accuracy nudge prompt on a Facebook-like social media platform (from Pennycook et al. (2021b)).

Nudging has emerged as one of the less-interventionist choices that policymakers have at their disposal (Thaler and Sunstein, 2009), the idea being that a gentle pointer towards desired behavior can be enough to influence outcomes in meaningful ways. The policy has been recently studied in the context of misinformation in field experiments such as Pennycook et al. (2021b, 2020b). While this has emerged as a popular choice given its uncontroversiality,

[11] See Hwang and Lee (2021), who study the efficacy of such a policy.

there are replication studies on accuracy nudging that question the statistical significance of this policy in reducing the influence of misinformation (Roozenbeek et al., 2021).

Provenance has also been discussed as an easily-implementable, less-invasive policy. The idea behind provenance is to facilitate users in their fact-checking process by providing them with more information about the original source of an article or post. For example, provenance may allow the user to see the chain of shares and easily trace the original context of a quote. We explore the benefits, and potential drawbacks, of such a policy in the model of Section 2.1.

1.3.5 Performance Targets

We consider a policy where the regulator imposes a *performance target* — a target that requires misinformation on the platform to be below a certain level.[12] This regulation transfers the burden of removing violating content (e.g., misinformation or hate speech) to the platform instead of an overseer (e.g., government agency), and has been proposed by Facebook in their own white paper (Bickert (2020)) as a preferred solution. A natural, but unrealistic, regulation is to require social media platforms to set the misinformation target at 0%. As Candogan and Drakopoulos (2020) identify, this is likely to decrease engagement on the platform . It also provides a plethora of perverse incentives; for example, it can shift the attention of the platform toward eradicating misinformation at the cost of neglecting other unmeasured/unregulated obligations, or it can lead to a narrower definition of what constitutes misinformation and make reporting it harder. Thus, while decreasing misinformation is beneficial, setting a performance target too low can have undesired effects. We will consider the problem of setting the optimal performance target for the platform.

1.3.6 Digital Advertising Tax

As we show in Chapter 5, platform business models that rely on digital advertising can be especially bad for consumer welfare. One of the most effective solutions we find for curbing platform incentives to design business models around digital ad targeting is . . . taxing digital

[12] In particular, Bickert (2020) proposes the following possible regulatory action:

> "Governments could also consider requiring companies to hit specific performance targets, such as decreasing the prevalence of content that violates a site's hate speech policies."

advertising revenue directly. Importantly, we identify some imperfect solutions, that under some conditions anti-trust interventions that breakup platforms or certain parts of the industry might improve consumer welfare. But as we discuss in that chapter, directly addressing the main issue, business model reliance on these targeted ads, can be especially effective.

Chapter 2

Misinformation: Strategic Models

This chapter focuses on the spread of misinformation in online social media networks. There are two classes of models I discuss in this thesis. The first class consists of *Bayesian models*, where agents make strategic decisions according to Bayesian reasoning and inference. In Sections 2.1 and 2.2, I propose two related Bayesian models of how misinformation might spread online, but where user sharing incentives are driven by different forces. To better understand why users share, I experimentally investigate misinformation sharing behavior in real social media environments in Section 2.2.4. The second class consists of *DeGroot models*, where agents are boundedly rational and act according to a behavioral heuristic. These are presented in Chapter 3.

2.1 A Model of Online Misinformation

In this section, we present a model based on Acemoglu et al. (2022b) of online content sharing where agents sequentially observe an article and decide whether to share it with others. This content may or may not contain misinformation. Agents gain utility from positive social media interactions but do not want to be called out for propagating misinformation. We characterize the (Bayesian-Nash) equilibria of this social media game and show sharing exhibits strategic complementarity. Our first main result establishes that the impact of homophily on content virality is non-monotone: homophily reduces the broader circulation of an article, but it creates echo chambers that impose less discipline on the sharing of low-reliability content. This insight underpins our second main result, which demonstrates that social media platforms interested in maximizing engagement tend to design their algorithms to create more homophilic communication patterns ("filter bubbles"). We show that platform incentives to amplify misinformation are particularly pronounced for low-reliability content likely to contain misinformation and when there is greater polarization and more divisive content. Finally, we discuss various regulatory solutions to such platform-manufactured misinformation.

2.1.1 Introduction

In this section, we develop a parsimonious model of online sharing behavior in the presence of misinformation, and as a first step, we focus on the behavior of fully Bayesian agents.[1] Our model is inhabited by a set of N agents. Each agent has a prior about the state of the world ("ideological bias"), and is connected to the rest of the users via a network, which is given by agents' friends and acquaintances, and is also shaped by the algorithms of the social media platform. A news article, defined by an underlying type (truthful or containing misinformation), a message (right-wing or left-wing), and a level of reliability (which determines the likelihood of misinformation), is then seeded at one of the agents. The message and the level of reliability of the article are common knowledge, while whether it is truthful or contains misinformation is unobserved, and agents form beliefs about this component.

Given these beliefs, the agent in question decides whether to ignore, dislike, or share the news article. If it is shared, the article moves from the agent sequentially to her connections on social media, who are then faced with the same choices. If the article is ignored or disliked, it does not get past the agent. We assume that agents receive utility when their shared content is re-shared and incur a cost when it is disliked. The former aspect captures the role of positive engagement in social media, while the latter represents the reputation loss from being called out for sharing content containing misinformation. Agents additionally receive utility from disliking (or calling out) items that they believe contain misinformation.

We characterize the Bayesian-Nash equilibria of this sequential game and prove that these equilibria always exist and are in cutoff strategies. In particular, our payoff structure implies that an individual will share any item that she believes is truthful with a high probability and

[1] Myopic reactions and biased behavior appear to play some role, for example, via the "confirmation bias" in social media behavior (see, e.g., Buchanan (2020) and Pennycook and Rand (2019)), but we believe that the Bayesian benchmark we construct already generates a number of empirically-relevant and rich results. We view incorporating realistic and relevant behavioral biases as a next step in this research agenda.

will dislike articles that she believes to contain falsehoods. Items with intermediate beliefs will be ignored. Beliefs about the truthfulness of articles are formed on the basis of the article's reliability and message, and agents' prior beliefs/ideology. Moreover, we establish that ours is a game of strategic complements: when others are more likely to share an item, each agent also becomes more likely to do so. As a result, we show that the set of equilibria forms a lattice, with well-defined most-sharing and least-sharing equilibria. All else equal, low-reliability articles are shared less, while articles that are "sensational" (either because they have provocative content or have broad appeal for other reasons) are shared more.

We present two main results. First, we study the implications of the (social media) network structure. We establish non-monotone comparative statics with respect to the degree of homophily (which determines how likely agents are to be connected to others who are ideologically similar to them). Low levels of homophily ensure that agents are likely to be exposed to cross-cutting content, including "counter-attitudinal articles" that advocate views opposed to theirs. This in turn ensures that misinformation is unlikely to survive for very long. Perhaps paradoxically, for high-reliability articles, an increase in homophily reduces content virality. This is because greater homophily makes it less likely that an article escapes a given community, reducing its circulation throughout the network.

More interestingly, when relevant news items have low reliability, homophily increases virality. This is because, countering the circulation effect, high homophily also creates a perverse incentive effect: knowing that shared articles will be seen by like-minded individuals, agents become more likely to share questionable content. Strategic complementarities amplify this effect, because when others are expected to share, the benefits from sharing are greater and being called out for spreading misinformation becomes less likely. It is particularly telling that homophily leads to the viral spread of low-reliability content, which are the ones more likely to contain misinformation.

We also show that political polarization and politically divisive articles are more likely to spread virally when they are low-reliability and the level of homophily is already high, generating an echo chamber-like social media environment. Strategic complementarities tend to amplify these pernicious effects of political polarization and divisive content as well.

Our second main result turns to social media platforms' algorithm design choices. We assume that platforms maximize engagement (in order to increase revenues from advertisements).

Under this assumption, we establish a striking result: when the relevant articles have low reliability, social media platforms design algorithms that increase homophily and create filter bubbles, propagating misinformation. Intuitively, high-reliability content tends to spread anyway because most users recognize it as such and share it, and low homophily contributes to its spread by increasing its circulation throughout the network. In contrast, low-reliability content will be ignored or disliked by agents who disagree with its message and believe it to contain misinformation. Engagement with low-reliability content can be boosted if the platform ensures that it remains among users ideologically aligned with its message, who would be willing to share it with like-minded others without fear of being called out by users with different ideologies. Hence, creating filter bubbles becomes an attractive strategy for engagement-maximizing platforms. It is particularly troublesome that such filter bubbles are created precisely when the relevant content is low-reliability and likely to contain misinformation.

If platform algorithms are propagating misinformation, can public policy counter and discourage this type of behavior?[2] The answer is yes, but with some caveats. In the last part of the paper, we discuss four different types of regulatory policies, and in each case, we show how they may reduce misinformation but also point out the possibility that, if they are not designed well, they can backfire and exacerbate the problem.

First, we look at potential censorship of articles identified by a regulator as likely containing misinformation. While censorship can help reduce the viral spread of misinformation, it also generates an "implied truth" effect (Pennycook et al. (2020a)) that contributes to the viral spread of questionable content that escapes censorship. Second, we discuss regulations that force platforms to reveal the provenance of articles, making it easier for users to identify falsehoods (e.g., claims originating from less reputable sources, such as InfoWars). Though generally useful and sometimes more powerful than censorship, provenance regulation can also backfire. This is for a related reason: this policy also creates an implied truth effect because individuals rely on other users' verification of the content before them. Third, we discuss "performance targets", where the regulator places limits on the amount of misinformation that circulates on the platform. Such targets tend to better align platform and regulator preferences, but unless

[2] As of August 2021, federal law protects social media platforms from being held responsible for content posted by its users (see Section 230 of the Communications Decency Act of 1996, discussed by `https://hbr.org/2021/08/its-time-to-update-section-230`).

they appropriately monitor and penalize platforms for violations, strict targets can exacerbate the spread of misinformation. Lastly, we show how direct regulation of platform algorithms can reduce misinformation, but also point out that the non-monotone effects of homophily imply that such regulations need to be finely calibrated.

Related Literature. Our paper builds on a large body of work on models of misinformation. In addition to the literature mentioned previously, several other papers in this literature are related to our findings.

Much previous work has focused on the susceptibility of boundedly-rational agents to engage with misinformation. In Acemoglu et al. (2010) and Acemoglu et al. (2013), the existence of persuasive agents can impede information aggregation and enable misinformed beliefs to survive, and sometimes even become dominant, in the population. In Mostagir et al. (2022) and Mostagir and Siderius (2022b), a strategic principal who wants to persuade agents of an incorrect belief can distort the learning process by leveraging social connections and echo chambers to propagate misinformation. Similarly, models of misinformation "contagion"—without Bayesian agents or strategic decisions—have been studied in Budak et al. (2011), Nguyen et al. (2012), and Törnberg (2018). Our contribution relative to this literature is the possibility that misinformation spreads because of the strategic interactions of Bayesian agents and is exacerbated by profit-maximizing platform algorithms.

There is a growing literature on information design by platforms, building for the most part on the concept of Bayesian persuasion (Kamenica and Gentzkow (2011) and Kamenica (2019)). Candogan and Drakopoulos (2020) study how a platform with private knowledge of content's accuracy should optimally signal to rational users whether to engage with it, while Chen and Papanastasiou (2021) and Keppo et al. (2019) consider more manipulative actions by platforms, including strategic seeding of information or "cheap talk" signals about quality. Also related are works on reputation and media bias. Motivated by the 2016 presidential election, Allcott and Gentzkow (2017) study the incentives of certain outlets to present misleading news, while Gentzkow and Shapiro (2006), Hsu et al. (2020) and Allon et al. (2021) explore other strategic reasons for media bias. Our paper contributes to this literature by highlighting the role of ideological leaning, strategic interactions, ideological homophily, and platform algorithms.

The most closely related work to ours is Papanastasiou (2020) who studies a model where agents hold heterogenous ideological beliefs and digest (and potentially share) a news article sequentially. Our work is different in three important dimensions. First, Papanastasiou (2020) focuses on costly inspection, which makes sharing decisions strategic substitutes, while our model generates strategic complementarities, because individuals care about further shares of the content they share.[3] All of our results and formal analysis turn on strategic complementarities. Second, and relatedly, echo chambers play no role in Papanastasiou (2020).[4] Third, our analysis of engagement-maximization by the platform and its implications for the spread of low-reliability content has no counterpart in Papanastasiou (2020) or any other work in this area we are aware of.[5]

The rest of the paper is organized as follows. The next section introduces our basic environment and describes the information structure and payoffs. Section 2.1.3 characterizes the (Bayesian-Nash) equilibria of this model and provides some basic comparative static results. Section 2.1.4 studies the effects of homophily by focusing on a special class of sharing networks that correspond to a set of "islands" of like-minded individuals who are less closely linked to those in other islands. Section 2.1.5 endogenizes the sharing network as a result of the algorithmic choices of the platform that aims to maximize engagement. Section 2.1.6 discusses a range of regulations aimed at containing misinformation. Section 2.1.7 concludes, while all proofs are provided in Appendix B.3.1.

2.1.2 Model

There is an underlying state of the world $\theta \in \{L, R\}$, for example, corresponding to whether the left-wing or the right-wing candidate is more qualified for political office. Agents have heterogeneous prior (ideological) beliefs about θ, and agent i's prior that $\theta = R$ is denoted by b_i with an *ex ante* distribution $H_i(\cdot)$, which may or may not be the same across agents.

Sharing Network. We assume there are N agents in the population, who share a news item according to a *sharing network* defined by a matrix $\mathbf{P}$ of link probabilities, with p_{ij} denoting

[3] Our reading of the evidence is that strategic complementarities are more relevant for social media behavior than strategic substitutabilities. For example, Eckles et al. (2016) find evidence that feedback or "encouragement" from peers about Facebook posts have contributed significantly to future behavior and posting. See also Taylor and Eckles (2018) and Aral and Dhillon (2018).

[4] As already noted, echo chambers appear central to the spread of misinformation in practice. See, for example, Lee et al. (2011), Törnberg (2018), Centola (2010), and Centola and Macy (2007)

[5] Papanastasiou (2020) also discusses platform incentives, but assumes that the platform is interested in limiting misinformation. Our reading of the evidence in this instance, too, favors our interpretation, where platforms such as Facebook are (or at the very least used to be before regulatory pressure mounted) fairly indifferent to the presence of misinformation but strongly prioritize engagement maximization.

the probability that agent i has a link to agent j. We define agent i's neighborhood $\mathcal{N}_i$ as the set of agents attached to her with an outgoing link, and denote her degree or the size of her neighborhood by $|\mathcal{N}_i|$. The sharing network reflects both an individual's social circle and the algorithms the platform uses for promoting shared content. The news item in question could be a news article or a post by one of the users, and throughout we refer to it as an "article".

Misinformation and News Generation. Each article has a three-dimensional type (r, m, ν). Here, $r \in [0, 1]$ indicates the *reliability* of the news, and $m \in \{L, R\}$ is the *message*, which corresponds to the article's viewpoint, for example, whether it argues for a left-wing or right-wing idea. Finally, ν is the article's *veracity*, which can either be $\mathcal{T}$, to indicate the article is truthful, or $\mathcal{M}$, to indicate the article contains *misinformation*.[6]

We assume that, at the beginning of the game, the type vector (r, ν, m) is drawn according to the following i.i.d. process:

(i) The article's reliability $r \in [0, 1]$ is drawn from a continuous distribution F with density f.

(ii) The veracity of the article is $\nu = \mathcal{T}$ (contains truthful content) with probability $\phi(r)$ or is $\nu = \mathcal{M}$ (contains misinformation) with probability $1 - \phi(r)$. We assume that ϕ is increasing and differentiable in r, and satisfies $\phi(0) = 0$ and $\phi(1) = 1$, so that the least reliable article always contains misinformation, and as the degree of reliability increases, the likelihood of misinformation monotonically declines and reaches zero.

(iii) If $\nu = \mathcal{T}$ (the article is truthful), then its message is generated as $m = \theta$ with probability $p > 1/2$. Conversely, if $\nu = \mathcal{M}$ (the article contains misinformation), then its message is generated as $m = \theta$ with probability $q \leq 1/2$ and is weakly anti-correlated with the truth.

While m and r are common knowledge (for example, the message m is directly observed and reliability depends on certain commonly-observed characteristics such as source and headline), the third dimension, ν, is unknown to all agents. We assume that agents update their

[6] Our focus in this paper is on misinformation, interpreted as items containing misleading information or arguments that can influence (a subset of) the public. Articles containing misinformation are in practice much more numerous than those that can be classified as "fake news", which explicitly propagate demonstrably false information (e.g., Egelhofer and Lecheler (2019), Allen et al. (2020), Guess et al. (2019), Grinberg et al. (2019)). For example, according to this definition a news item that favorably describes a report denying climate change, without putting this in the context of hundreds of other reports reaching the opposite conclusion or mentioning the criticisms that it has received from experts, contains misinformation.

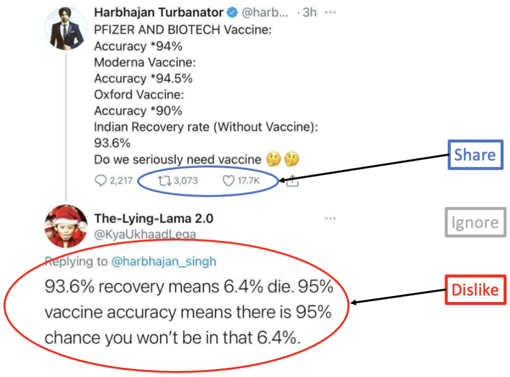
beliefs about ν using Bayes' rule given beliefs about the underlying state θ and the observables (r, m) of the article.

Social Media Behavior. Time is discrete $t = 1, 2, \ldots$. Upon receipt of the article, an agent i can take one of three actions $a_i \in \{\mathcal{S}, \mathcal{I}, \mathcal{D}\}$, as described below:

(i) Share ($\mathcal{S}$): The agent decides to *share* the article and passes it onto others after her.

(ii) Ignore ($\mathcal{I}$): The agent decides to *ignore* the article and does not engage with it.

(iii) Dislike ($\mathcal{D}$): The agent decides to *dislike* the article, which means expressing disagreement or contempt for the content contained in it.


Harbhajan Turbanator @harb... · 3h
PFIZER AND BIOTECH Vaccine:
Accuracy *94%
Moderna Vaccine:
Accuracy *94.5%
Oxford Vaccine:
Accuracy *90%
Indian Recovery rate (Without Vaccine):
93.6%
Do we seriously need vaccine 🤔🤔

The-Lying-Lama 2.0
@KyaUkhaadLega
Replying to @harbhajan_singh
93.6% recovery means 6.4% die. 95% vaccine accuracy means there is 95% chance you won't be in that 6.4%.

Figure 2-1. Sample Tweet.

The three possible actions are depicted in Figure 2-1 using Twitter as a sample social media platform. A given user sees the article and decides how to respond to it. She can (i) share it ($\mathcal{S}$), which actively puts it on other social media news feeds; (ii) ignore it ($\mathcal{I}$), where the user simply scrolls past the article; or (iii) actively dislike it ($\mathcal{D}$), expressing derision for the content.

At time $t = 1$, we assume that some initial seed agent i^* first engages with the article. If the article is shared by agent i, it is passed to all $j \in \mathcal{N}_i$. In contrast, following ignore or dislike, the article does not propagate past agent i.

Payoffs. Let us define shares after i as $S_i = |\{j \in \mathcal{N}_i \ : \ a_j = \mathcal{S}\}|$ and dislikes after i as $D_i = |\{j \in \mathcal{N}_i \ : \ a_j = \mathcal{D}\}|$. Agent i's utility can then be written as

$$U_i = \begin{cases} 0, & \text{if } a_i = \mathcal{I} \\ \tilde{u}\mathbf{1}_{\nu=\mathcal{M}} - \tilde{c}, & \text{if } a_i = \mathcal{D} \\ u\mathbf{1}_{\nu=\mathcal{T}} - c\mathbf{1}_{\nu=\mathcal{M}} + \kappa S_i - dD_i, & \text{if } a_i = \mathcal{S} \end{cases} \tag{2.1}$$

where $\mathbf{1}$ is the indicator function (equal to 1 if true and 0 otherwise). Here, $\tilde{u}, \tilde{c}, u, c, \kappa$ and d are strictly positive parameters, which we discuss below.

(i) We normalize payoffs following ignore, $\mathcal{I}$, to $U_i = 0$.

(ii) Payoffs from dislike, $\mathcal{D}$, depend on whether the article contains misinformation. We assume, in particular, that disliking has a cost of $\tilde{c} > 0$, regardless of whether the article is truthful (because of, say, the effort required to actively call out misinformation). In addition, disliking an article containing misinformation has a benefit of $\tilde{u} > \tilde{c}$, because individuals like calling out misleading articles. This formulation implies that disliking is never preferred to ignoring for an article that is truthful with probability 1, and is always preferred to ignoring for an article that contains misinformation with probability 1.

(iii) Following a decision to share, $\mathcal{S}$, an agent receives utility from two sources. First, agents receive utility from sharing truthful content, but incur a cost from sharing misinformation. This explains the first component of utility following $\mathcal{S}$, $U_i^{(1)} = u\mathbf{1}_{\nu=\mathcal{T}} - c\mathbf{1}_{\nu=\mathcal{M}}$. Second, agents enjoy positive feedback from their peers (such as likes, or in our setting re-shares), but are negatively affected by dislikes. This is captured by the second component of utility $U_i^{(2)} = \kappa S_i - dD_i$.[7] In this formulation, the parameter κ captures the importance of "popularity" for the agent's sharing decision, while d represents the extent to which she cares about negative reactions. In Appendix A.1.2 we provide a simple microfoundation for disutility from negative reactions based on reputational concerns.

Information Structure and Solution Concept. Agents are not aware of, and have uniform prior over, when the article was first introduced onto social media, the prior sharing process, and the structure of the social network (though the link matrix $\mathbf{P}$ is common knowledge).[8] Moreover, while any agent i knows the distribution $\{H_i\}_{i=1}^N$ of beliefs in the population, she does not know any agent j's belief (ideology) b_j. We focus on Bayesian-Nash equilibria, and refer to these as "equilibria" for short.

[7] Equivalently, the terms κS_i and dD_i could be replaced by arbitrary functions $\varphi_S(\kappa, S_i)$ and $\varphi_D(d, D_i)$ that satisfy $\varphi_S(0,\cdot) = \varphi_S(\cdot, 0) = \varphi_D(0,\cdot) = \varphi_D(\cdot,0) = 0$ and have (weakly) increasing differences. This generalization captures a broad range of peer feedbacks based on sharing different types of content, beyond the additive structure we adopted for notational simplicity in the text.

[8] Hence, agents do not know the exact interactions and sharing patterns outside their neighborhood, which is consistent with the evidence in Breza et al. (2018). That being said, because the equilibrium sharing process is Markovian, this assumption can be relaxed by replacing $\mathbf{P}$ with the adjacency matrix.

To eliminate trivial and unrealistic equilibria, we assume that the sensationalism of an article is upper bounded by $\bar{\kappa} = (c\tilde{c} - u(\tilde{u} - \tilde{c}))/(\tilde{u}N)$. This assumption guarantees that there is never an equilibrium where *every* agent *always* shares *all* articles. It also eliminates equilibria where agents may share and dislike, but never ignore.

Discussion—The basic assumptions introduced above are consistent with salient patterns of behavior and information structure in social media. First, as documented in studies such as Pennycook et al. (2021b), users want to share content they believe to be truthful and not contain misinformation. Second, while users derive value from peer encouragement and re-shares on social media (Eckles et al. (2016)), they also suffer reputational costs when they get called out for sharing misinformation (see, for example, evidence from Facebook in Altay et al. (2020)). Finally, social media users often engage in criticisms of available content and inform others about misinformation (see, for example, Kim et al. (2020) for evidence in the context of 2018 midterm elections).

2.1.3 Equilibria in General Networks

In this section, we characterize the structure of equilibria for any sharing network structure P and provide various comparative statics. Without loss of generality (and ease of exposition), we fix the article's message as $m = R$ for the remainder of the paper.[9]

Cutoff Strategies and Strategic Complementarities

When agent i receives an article with reliability r and message $m = R$, she updates her (*ex post)* belief, π_i, that the article is truthful according to Bayes' rule:

$$\pi_i = \frac{(pb_i + (1-p)(1-b_i))\phi(r)}{(qb_i + (1-q)(1-b_i))(1-\phi(r)) + (pb_i + (1-p)(1-b_i))\phi(r)}. \tag{2.2}$$

Clearly, π_i is increasing in b_i since an agent is more likely to believe in an article's veracity when its message agrees with her prior. Moreover, π_i is increasing in r, as the agent updates more on the basis of more reliable articles.

[9] To see that this is without loss of generality, observe that the analysis applies identically with an $m = L$ message but with complementary priors $b_i' = 1 - b_i$.

We can also see that the payoff to sharing ($\mathcal{S}$) increases in π_i, since the first component of utility, $U_i^{(1)}$, is increasing in π_i (as the individual would like to share truthful articles), while $U_i^{(2)}$ is independent of π_i. With a similar reasoning, the payoff to disliking ($\mathcal{D}$) is decreasing in π_i, whereas the payoff to ignoring ($\mathcal{I}$) is independent of π_i. This monotone behavior of payoffs will lead to best-response decision rules for agents that take the form of cutoff strategies, as we explain next.

We say that agent i employs a *cutoff strategy* if there exists $b_i^*(r)$ and $b_i^{**}(r)$ such that agent i chooses $\mathcal{S}$ when $b_i > b_i^{**}(r)$, chooses $\mathcal{I}$ when $b_i^*(r) < b_i < b_i^{**}(r)$, and chooses $\mathcal{D}$ when $b_i < b_i^*(r)$. Cutoff strategies in our context imply that agents who strongly agree with an article tend to share it, agents who strongly disagree with it tend to choose dislike, and those with intermediate beliefs typically ignore the article.

We will see in the next theorem that all equilibria are in cutoff strategies. This means, in particular, that an equilibrium can be summarized by cutoff vectors $(\mathbf{b}^*, \mathbf{b}^{**}) = (b_1^*, b_1^{**}, \ldots, b_N^*, b_N^{**})$. Furthermore, these cutoffs $b_i^*(r)$ and $b_i^{**}(r)$ will both be decreasing in r, so that as reliability increases, an article becomes more likely to be shared and less likely to be disliked.

We can also note that our social media game exhibits *strategic complementarities*. To see this, observe that when others share more—meaning that b_i^{**} (weakly) decreases for all i—the second component of utility, $U_j^{(2)}$, increases for each agent j, and this raises the overall utility of sharing and encourages more sharing. Similarly, when others reduce their likelihood of disliking, meaning that now b_i^* (weakly) decreases for all i, this reduces the likely cost of sharing misinformation by mistake, also raising $U_j^{(2)}$. Strategic complementarities capture an important dimension of social media interactions—utility feedback from others' behavior tends to encourage agents to cohere with those behaviors.

Equilibrium Structure. The next theorem shows that an equilibrium always exists and is in cutoff strategies. At the same time, strategic complementarities ensure that there is a well-defined structure to the set of equilibria. To make this more concrete, we say that an equilibrium $(\mathbf{b}^*, \mathbf{b}^{**})$ has *uniformly more sharing* than other $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$ if $\mathbf{b}^* \preceq \hat{\mathbf{b}}^*$and $\mathbf{b}^{**} \preceq \hat{\mathbf{b}}^{**}$ (where $\preceq$ is the component-wise order). This, in particular, means that all the thresholds for each agent is (weakly) lower in the former equilibrium (recall that lower thresholds mean more sharing).

Theorem 2.1.1.

(i) There exists a Bayesian-Nash equilibrium;

(ii) All equilibria are in cutoff strategies;

(iii) The set of cutoffs $(\mathbf{b}^{*}, \mathbf{b}^{**})$ *forms a lattice, and thus there exists a least-sharing and most-sharing equilibrium.*

The structure of equilibria characterized in Theorem 2.1.1 facilitates our analysis, enabling us to focus on two sets of thresholds, $(\mathbf{b}^{*}, \mathbf{b}^{**})$, which are themselves monotone in the reliability of the article in question. Because of strategic complementarities, there can be multiple equilibria: when others are choosing to share an article with middling reliability, this further encourages sharing because one's own post will circulate more, increasing the utility from sharing. Conversely, if the same article with middling reliability is not shared by others, the payoff to sharing is reduced, while the cost of being found out to have circulated misinformation remains constant. This then discourages sharing.

Theorem 2.1.1 also shows that, despite this multiplicity, there are two focal equilibria on which we can concentrate: the equilibrium with the smallest vector of cutoffs (*most-sharing equilibrium*) and the equilibrium with the largest vector of cutoffs (*least-sharing equilibrium*).

Finally, the theorem's characterization provides an explicit measure of the amount of sharing. Recall that agent i's prior belief b_i is drawn *ex ante* from the distribution H_i. Hence, in an equilibrium with cutoff b_i^{**} for agent i, the *ex ante* likelihood that this agent will share is $1 - H_i(b_i^{**})$. Therefore, the most sharing equilibrium, which has the smallest equilibrium b_i^{**} for all i, has the highest likelihood that any agent i will share the article in question.

Content Virality and Comparative Statics (for Fixed $\mathbf{P}$)

In this subsection we provide comparative statics for the most and least sharing equilibria as we change the parameters of the social media game, but holding the sharing network $\mathbf{P}$ fixed. We discuss comparative statics with respect to the network in the next section. Toward this goal, we define the notion of *content virality*, which captures the expected spread of an article in the sharing network.

Content Virality. Formally, we define content virality as follows. We suppose that an article is seeded at some agent i^{*} at $t = 1$. We then define $\mathbf{S}_{i^{*}}$ as the (random) proportion of the population that shares when agent i^* is the seed agent in the most-sharing equilibrium σ. We say σ_1 has more *content virality* than σ_2 if $\max_{i^*} \mathbb{E}_{\sigma_1}[\mathbf{S}_{i^*}] \geq \max_{i^*} \mathbb{E}_{\sigma_2}[\mathbf{S}_{i^*}]$. In words, our notion of content virality compares the spread of an article provided that it starts from the seed that is most favorable to its ultimate circulation. The reason we start from the most favorable seed is that, as we will see in Section 2.1.5, social media platforms have an incentive to implement sharing algorithms that place articles in such favorable seeds. For future reference, we also note that content virality is also the same as expected overall engagement with an article, conditional on favorable seeding.

Quantity of Misinformation, Sensationalism, and Reputation. The next proposition shows how the quantity of misinformation, sensationalism, and reputational concerns affect content virality. We define less misinformation as a shift of the function ϕ to some $\phi' \geq \phi$ (pointwise). The parameter κ captures how *sensational* the article is: higher κ implies that agents receive greater value from future shares, because these shares are associated with others paying more attention or perhaps being entertained more by the relevant posts. This greater utility is independent of the content's veracity. We think of κ varying at the level of articles, so that some articles will be more sensational than others. Finally, the parameter d proxies for for the importance of *reputational concerns*. Higher d means that dislikes are more damaging, which corresponds to the agent being more concerned about receiving many dislikes. We think of d as varying at the level of communities (certain communities of users, for example, academics, may have more reputational concerns).

Proposition 2.1.1. *Less misinformation, higher sensationalism, and weaker reputational concerns lead to greater content virality.*

These results are intuitive and immediate.[10] Holding constant the reliability of the article, less misinformation reduces the cost associated with sharing, triggering more aggressive sharing by all agents. This prediction is consistent with Pennycook and Rand (2019) who show low-reliability content (e.g., Breitbart or Infowars) is not typically shared by attentive social media users, regardless of partisanship. This proposition also clarifies that viral spread of misinformation is not a mechanical effect in our model: if anything, less reliable articles that

[10] In fact, the claim in Proposition 2.1.1 can be strengthened to the notion of *uniformly more sharing* where *all* agents in the sharing network share with strictly higher probability, which immediately implies higher content virality. We focus on content of virality, both because it is simpler and also because the comparative static results with respect to homophily in the next section do not always lead to uniformly more or less sharing.

are more likely to contain misinformation are less likely to become viral. We will see that other aspects of social media interactions, in particular the topology of the sharing network, are often responsible for viral spread of misinformation.

The comparative static with respect to sensationalism coheres with the patterns documented in Duffy et al. (2020), suggesting that social media participants often share a story that is "too good not to share", and do so even when they realize it is also "too good to be true". The link between reputational concerns and misinformation is also consistent with the evidence that in settings where reputation matters misinformation is less likely (Altay et al. (2020)), and when such reputational concerns are missing, even calling out individuals sharing misinformation is fairly ineffective (Mosleh et al. (2021a)).

Finally, this proposition provides a possible pathway for low-reliability content to become viral. Vosoughi et al. (2018) argued that misinformation spreads farther, faster, deeper and more broadly than truthful news on social media. This evidence was criticized by Grinberg et al. (2019) who showed that, once the effects of sensational news items is controlled for, misinformation does not spread farther (or faster) than truthful content. Proposition 2.1.1 provides a rationalization of these patterns. All else equal, misinformation does not spread faster than truthful content as in Grinberg et al. (2019). However, because, as observed in Molina et al. (2021) and Kozyreva et al. (2020), sensational content is often low-reliability, these two propositions together imply that misinformation may be more likely to become viral. Whether this happens or not depends on the boost from sensationalism. When this is limited, low-reliability articles containing misinformation spread less because of concerns of users that others will call them out for sharing this content. But when this sensationalism boost is high, misinformation can become viral. Additionally, the strategic complementarity in sharing decisions implies that sufficiently sensational misinformation can become viral, because once an individual thinks others are going to share this sensational item, she becomes much more likely to share herself, even if she has doubts about its veracity.

2.1.4 Island Networks and the Implications of Homophily

In this section, we present comparative static results with respect to the sharing network P. Throughout this section, we take the sharing network as given, and then return to how it is shaped by the algorithms of social media platforms in Section 2.1.5.

The focus on comparative statics with respect to the sharing network necessitates two modifications from our analysis so far. First, we restrict attention to *island networks* (or equivalently, the stochastic block model), which are lower-dimensional than general networks we have allowed so far. Namely, in an island network, agents are partitioned into k blocks of size $N_1, N_2, \ldots, N_k$, called *islands* each with some constant (but not necessarily equal) share of the population N. Each agent i has a type $\ell_i \in \{1, \ldots, k\}$ corresponding to which block (or "island") she is in. Link probabilities are then given as:

$$p_{ij} = \begin{cases} p_s, \text{ if } \ell_i = \ell_j \\ p_d, \text{ if } \ell_i \neq \ell_j \end{cases}$$

where $p_s \geq p_d$. Without loss, we assume each of the islands is weakly connected.

Second, we assume the prior distribution for agents on the same island ℓ is the same, and is denoted by H_ℓ. We also assume that islands are ranked according to their belief distributions. In particular, each island ℓ has distribution H_ℓ with support on $[b^{(\ell)}, b^{(\ell+1)}]$, where $1 \geq b^{(1)} > b^{(2)} > \ldots > b^{(k)} > b^{(k+1)} \geq 0$.[11] This implies that lower-indexed islands have stronger right-wing beliefs.

An important advantage of island networks, in addition to their lower-dimensional representation, is that, combined with this ranking assumption, they enable us to model the degree of *homophily*—the extent to which an individual interacts with others that have common characteristics as herself. Common characteristics for us are those that are relevant for prior beliefs, and therefore, by construction, individuals have more in common with those on the same island as themselves. As a result, homophily will be higher when most links are within islands and links between islands are sparse (high p_s and low p_d).[12]

More formally, we say that an island network with (p_s, p_d) has *more homophily* than an island network with (p'_s, p'_d) if all agents have the same expected degree under both, but where $p_s > p'_s$ and $p_d < p'_d$.[13] From Theorem 2.1.1, we know that the equilibrium is in cutoff strategies

[11] This assumption is adopted for simplicity. Our results generalize if we instead assume that these distributions are ranked in terms of first-order stochastic dominance: $H_1 \succeq_{FOSD} H_2 \succeq_{FOSD} \cdots \succeq_{FOSD} H_k$. However, this generalization requires considerably more formalism and notation, motivating our focus on disjoint supports.

[12] The homophilic structure and greater congruence of beliefs within islands are consistent with the evidence presented in Bakshy et al. (2015): "friend networks" on Facebook are ideologically segregated, with the median share of friends from the opposing ideology around only 20%. Mosleh et al. (2021b) provides evidence of similar homophily on Twitter.

[13] Because network density raises connectivity and can directly increase virality, we hold network density fixed

of the form $(\mathbf{b}^*, \mathbf{b}^{**}) \equiv (b_1^*, b_1^{**}, \ldots, b_N^*, b_N^{**})$. However, because there is symmetry within islands, equilibria now take a simpler, "semi-symmetric" form as shown in the next lemma.

Lemma 2.1.1. *All equilibria are semi-symmetric: for every equilibrium, there exist* $\{(b_\ell^*, b_\ell^{**})\}_{\ell=1}^k$ *such that* $b_i^* = b_{\ell_i}^*$ *and* $b_i^{**} = b_{\ell_i}^{**}$ *for all agents* i *in island* ℓ*.*

The simplification established in Lemma 2.1.1 will allow us to work with a lower dimensional cutoff vector (just two cutoffs for each island).

Comparative Statics: Homophily

The next theorem, characterizing the effects of homophily on the spread of misinformation, is our first main result:

Theorem 2.1.2. *There exist* $0 < \underline{r} < \bar{r} < 1$ *such that:*

(a) If $r < \underline{r}$*, an* increase *in homophily increases the virality of content.*

(b) If $r > \bar{r}$*, a* decrease *in homophily increases the virality of content.*

Theorem 2.1.2 shows how low-reliability content can spread virally in networks with high homophily. Intuitively, when content comes from a low-reliability source, only agents who (strongly) agree with the article's message share it. However, as homophily increases, users know that they will mostly share with other like-minded people, who will also be inclined to share this content. This creates a type of echo chamber: the likelihood of being called out for spreading misinformation is now lower, making users "less disciplined" or more likely to share lower-reliability content. Strategic complementarities then extend these incentives throughout the network. In this way, homophily leads to the viral spread of low-reliability articles that likely contain misinformation.

However, Theorem 2.1.2 shows that homophily can have non-monotone effects. This is because greater homophily also keeps an article circulating among the same group of like-minded users and reduces the likelihood that it will reach other communities. Theorem 2.1.2(a) establishes that the first effect of homophily, working through incentives to share low-reliability content, is more powerful than the second, "circulation effect", when we focus on particularly

in order to isolate the effects from homophily.

low-reliable content (with $r < \underline{r}$). This implies, in particular, that homophily's impact is to increase the virality of especially low-quality content, which is of course relevant for public policy (as we discuss in Section 2.1.6).

The results in Theorem 2.1.2 are in line with recent evidence highlighting the importance of echo chambers for the spread of misinformation. Törnberg (2018) and Vicario et al. (2016), among others, show that homophily in sharing behavior propagates ideologically-congruent ideas, with little incentive to question the veracity of this information, while Quattrociocchi et al. (2016) document how echo chambers on Facebook fuel conspiracy theories and the popularization of incorrect scientific ideas, for example, on vaccines. Levy (2021) provides evidence that "filter bubbles" generated by Facebook's algorithms are an important source of propagation of misinformation.

Comparative Statics: Divisive Content and Belief Polarization

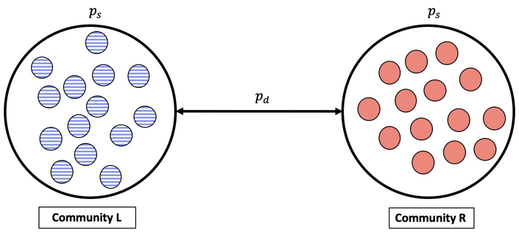


Figure 2-2. Two-Island Model.

In this subsection, we provide an additional comparative static with respect to polarization. For this result, we focus on the case of just two islands, a left-wing and a right-wing one with prior distributions H_L and H_R, respectively, as pictured in Figure 2-2. Moreover, we suppose that there is disjoint support of prior beliefs across communities. Formally, we assume H_R has support on $[\underline{b}_R, \bar{b}_R]$ and H_L has support on $[\underline{b}_L, \bar{b}_L]$, with $\bar{b}_L < 1/2 < \underline{b}_R$.

We say content with parameters (p', q') is *more divisive* than content with parameters (p, q) if $p \geq p'$ and $q \leq q'$. Divisive content has a message that is more tethered to the true state θ when it is truthful (and more likely to argue against θ if it is misinformation). In our case, we think of state θ as related to political ideology. Therefore, non-political content, such as wedding photos or cat videos, has little divisiveness relative to more political ones, such as "Obama Signs Executive Order Banning The Pledge of Allegiance in Schools Nationwide" (Fourney et al.

(2017)). Note also that, from equation (2.2), prior beliefs about θ matter more for updating and the assessment of an article's veracity when content is more divisive.

We say H_2 is *more polarized* than H_1 if it satisfies the following single crossing property: $H_2^{-1}(\alpha) - H_1^{-1}(\alpha)$ is a nondecreasing function in α, crossing zero at $\alpha^* = 1/2$ with $H_1(1/2) = H_2(1/2) = 1/2$. An increase in polarization results in a "stretching" of the belief distribution around the most moderate user (i.e., $b = 1/2$) while preserving an equal distribution of left-wing and right-wing agents (meaning that $H(1/2) = H(1/2) = 1/2$, which is applied in the island model to the average distribution of beliefs, $H = \frac{1}{N}\sum_{\ell=1}^{N} N_\ell H_\ell$). The available evidence indicates that the US public has become more polarized (see Pew Research Center (2014) and Abramowitz (2010)), and an important question of debate has been whether this polarization has fueled the spread of misinformation on social media.

The next result studies how political divisiveness and polarization impact social media behavior and the spread of misinformation, as a function of the homophily in the sharing network.

Proposition 2.1.2. *There exist* $r^* \in (0,1)$ *and* $p^* \in (0,1)$ *such that:*

(a) If $r < r^*$ and $p_s/p_d > p^*$, then an *increase* in divisiveness or an increase in polarization leads to greater content virality.

(b) If $r > r^*$ and $p_s/p_d < p^*$, then a *decrease* in divisiveness or a decrease in polarization leads to greater content virality.

Proposition 2.1.2 is complementary to Theorem 2.1.2. When the content in question has high reliability ($r > r^*$) and homophily is limited, more divisive content or greater polarization tends to reduce content virality, because in a well-connected, non-homophilic network, controversial articles will solicit a wide range of reactions, disciplining those tempted to share misinformation. In contrast, when the article in question has low reliability and there is significant homophily, there are again echo chamber-like effects. More divisive content generates more divergent behavior from individuals with different ideologies, and greater polarization means there are sharper differences in terms of these ideologies. As a result, echo chambers matter especially for divisive content and in the presence of polarization. Strategic complementarities once again amplify this effect, as users recognize that others in their community will tend to share divisive content, and this makes them even more willing to share.

It is notable that Proposition 2.1.2, like Theorem 2.1.2, implies that greater divisiveness and polarization increase the virality of especially low-reliability content, which is most likely to contain misinformation. These two results together thus imply that echo chambers, greater political polarization and divisive content all exacerbate the circulation of misinformation on social media.

2.1.5 Platform Design and Filter Bubbles

We now turn to our second main result: how platform behavior affects misinformation. Consider a collection of social media users with beliefs distributed according to a distribution H. The platform can identify communities of users according to prior ideological beliefs, for example, based on content previously shared or affiliations with ideological groups. In particular, each user is binned into one of k communities, with each community ℓ having a belief distribution H_ℓ with support over $[b_\ell, b_{\ell+1}]$, and where $1 \geq b^{(1)} > b^{(2)} > \cdots > b^{(k)} > b^{(k+1)} \geq 0$ (with at least one left-wing and one right-wing community). The size of these bins depends on the platform's microtargeting technology at identifying users' ideological beliefs (see, for example, Papakyriakopoulos et al. (2018)). Formally, we let $\varepsilon \equiv \max_\ell(b_{\ell+1} - b_\ell)$, with the interpretation that lower values of ε correspond to better platform technology for identifying ideology.

The platform's objective is to maximize user engagement, which is equivalent to maximizing content virality (see the definition of content virality in Section 3.3.4).[14] The platform does not directly care about whether the content users are engaging with is truthful or contains misinformation.

The platform chooses how content is shared across users. That is, for each article, the platform not only picks the seed agent at $t = 1$ to whom it recommends this article, but also

[14] This objective is rooted in the fact that social media sites, like Facebook, primarily rely on advertising revenue, which becomes more valuable as users increase their activity on the site. For example, 85% of Facebook's total revenue in 2011 was from advertising, and from 2017-2019, around 98% was (see Andrews (2012) and `https://www.nasdaq.com/articles/what-facebooks-revenue-breakdown-2019-03-28-0`).

Strictly speaking, we are modeling social media platform objectives before the more recent public backlash over misinformation. If the platform faces potential penalties from public backlash or regulators for spreading misinformation, its objective function will change, as we explore in greater detail in Section 2.1.6.

chooses the sharing network—the matrix of link probabilities $\mathbf{P}$.[15] The platform's choice of $\mathbf{P}$ can be interpreted as its "algorithm" to determine how users are exposed to content circulating in the social media site. This algorithm choice is assumed to be common knowledge.

Optimality of Island Networks and Filter Bubbles

We remind the reader that the island networks of Section 2.1.4 are parameterized by three components: the within-island link probability (p_s), the across-island link probability (p_d), and the number of islands (k). As special cases, we have (i) an island model that has *maximal homophily*, where $p_s > 0$ but $p_d = 0$ (and thus there is extreme ideological segregation on the network); and (ii) an island model with *maximal connectivity*, where $p_s = p_d$ (and there is minimal homophily and no segregation by ideology). We next show that although the platform is allowed to design any sharing network $\mathbf{P}$, its profit-maximizing choice is within the class of island networks.

Theorem 2.1.3. *There exists $\bar{\varepsilon} > 0$ such that if $\varepsilon < \bar{\varepsilon}$, the platform's profit-maximizing sharing network is determined by a reliability threshold $r_P \in (0, 1)$ such that:*

(i) If $r < r_P$, the platform's profit-maximizing sharing network has maximal homophily.

(ii) If $r > r_P$, the platform's profit-maximizing sharing network has maximal connectivity.

Part (i) of the theorem shows that when articles are mostly unreliable—and likely to contain misinformation—the platform creates an extreme filter by designing its algorithms to achieve a sharing network with the greatest homophily. In contrast, part (ii) demonstrates that when articles have higher reliability, the platform refrains from introducing algorithmic homophily. This result highlights an important channel by which misinformation spreads: it is precisely when articles are likely to contain misinformation that the platform seeks to maximize engagement by creating (endogenous) echo chambers, or filter bubbles, where these articles spread virally within like-minded communities. Put differently, with low reliability content, neither the platform nor the users are disciplined about sharing misinformation, and so these news items spread virtually uninhibited.

[15] Facebook's algorithms may induce different sharing networks depending on features of the article, such as whether it contains cat videos, wedding photos, or political content. See `https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/`.

We also remark that the threshold r_P parameterizes the extent of filter bubbles on the platform. Specifically, the most extreme left-wing agent will be exposed to all $m = L$ articles, regardless of their reliability, but only see right-wing articles with reliability $r \geq r_P$. Similarly, the most extreme right-wing agent will see all $m = R$ articles, but only $m = L$ articles with $r \geq r_P$.

It is further worth noting that this theorem builds on but also significantly strengthens Theorem 2.1.2. In Theorem 2.1.2, the effects of homophily are non-monotone and are ambiguous when an article is neither very low reliability nor very high reliability ($r \in (\underline{r}, \bar{r})$). In contrast, Theorem 2.1.3 gives a sharp characterization of the platform's algorithm: when $r > r_P$, the platform goes for maximal connectivity, and when $r < r_P$, it chooses maximal homophily, and in this case, misinformation spreads virally precisely because of the echo chambers that the platform has manufactured. In both cases, the island structure of the network considered in Section 2.1.4 arises endogenously as the profit-maximizing sharing network for the platform.

In addition, Theorem 2.1.3 shows that when a platform can shape the network topology through its recommendation algorithm, the echo chamber effect that arises from Theorem 2.1.2(a) is exactly what fuels misinformation (whereas in Theorem 2.1.2(b), echo chambers were harmless). This observation generalizes to cases where the platform has less precise microtargeting technology, albeit in a less sharp way than the result presented in Theorem 2.1.3.[16]

Remark—In Theorem 2.1.3, we assume the platform can select any network **P** it desires through its recommendation algorithm. This is without loss of generality. If we assume the social network originally begins as an arbitrary island network in Section 2.1.4, and the platform can hide and amplify content across different links, the same result readily follows.

[16] When $\varepsilon > \bar{\varepsilon}$, the platform still induces echo chamber-like environments to generate viral spread of low-reliability content, even if its choice does not take the form of an island network. For instance, if there are only three communities (with broad ideology spectra), a misinformation right-wing article can still spread virally via the usage of filter bubbles. However, the platform may now prefer a sharing network that lies outside the class of island networks considered in Section 2.1.4. Specifically, user engagement may be maximized if the left-wing community is completely disconnected from all other communities, but the moderate (middle) community has sparse connections to the right-wing community. This facilitates a strong echo chamber within the right-wing community but also allows the article to spread to the more moderate community (while receiving no discipline from the left-wing community).

Comparative Statics on r_P: Divisiveness and Polarization

In Theorem 2.1.3, the threshold r_P fully summarizes the extent to which misinformation will spread virally on social media.

We next perform comparative statics for this threshold to understand the conditions under which the platform will create a filter bubble and propagate misinformation.

Proposition 2.1.3. *The reliability threshold r_P increases as message divisiveness and/or belief polarization increases.*

Proposition 2.1.3 mimics the conclusions of Proposition 2.1.2(a). As divisiveness or polarization increases, content is consumed more aggressively within echo chambers and scrutinized more aggressively outside of them. Under these conditions, filter bubbles become more advantageous to the platform, especially when the relevant content has low reliability. This is because communities with more extreme beliefs now feel more strongly about news in general, rarely second-guess politically-congruent news, but often doubt and dislike counter-attitudinal news. As a result, low-reliability content spreads virally inside the platform's filter bubbles. In contrast, outside of the filter bubble, this content would have been quickly disliked and stopped—which is the reason why the platform favors algorithms that induce such filter bubbles.

Proposition 2.1.3 also provides a possible (albeit of course speculative) interpretation for why accelerating political polarization and identity politics in the last two decades may have come with more aggressive filter bubble algorithms from social media sites (Apprich et al. (2018)). As the recent documentary *The Social Dilemma* puts it: "The way to think about it is as 2.5 billion Truman Shows. Each person has their own reality with their own facts. Over time you have the false sense that everyone agrees with you because everyone in your news feed sounds just like you." Tellingly in this context, while Facebook cracked down on misinformation prior to the 2020 election in part due to political pressure, its algorithms have resumed promotion of misinformation in November and December of 2020: "...the measures [Facebook] could take to limit harmful content on the platform might also limit its growth: In experiments Facebook conducted last month, posts users regarded in surveys as 'bad for the world' tended to have a greater reach—and algorithmic changes that reduced the visibility of those posts also reduced users' engagement with the platform...".[17]

[17] See *Vanity Fair*:

2.1.6 Regulation

Our analysis so far raises the natural question of what types of regulations might counter the viral spread of misinformation and platform choices leading to excessive ideological homophily. We now briefly discuss four distinct types of regulations that have been discussed in this context: (1) *censorship* or tagging of misinformation; (2) regulations that force platforms to reveal articles' *provenance*; (3) *performance targets* that require the platform to keep misinformation below a given threshold; and, (4) *network regulations*, restricting the extent of ideological homophily or segregation introduced by platform algorithms intended to maximize engagement.

We consider the effects of these policies when the platform can optimally choose the sharing network in response to the public policy. For simplicity, we suppose the regulator's objective is to decrease the virality of articles containing misinformation on the platform. We say a policy is more *effective* than another policy (or no policy) if it reduces the virality of misinformation (and is *most* effective if more effective than any other feasible policy). Throughout, we fix the reliability of the article and assume the most-sharing equilibrium before the regulation involves some agents sharing and some agents not sharing (i.e., $\mathbf{b}^* \neq \mathbf{0}$ and $\mathbf{b}^{**} \neq \mathbf{1}$), allowing for the possibility that regulation might backfire and increase the virality of low-reliability content, or potentially help by reducing the virality of such content likely to contain misinformation.

Censorship

We first consider a policy where the regulator can censor misinformation that appears on the platform (also known as "content moderation").[18] Formally, we model this as the regulator being able to adopt a policy that removes at most $\delta \in (0, 1)$ fraction of the content containing misinformation (with each piece of misinformation removed with probability δ).[19] In other words, the regulator selects $\delta^* \leq \delta$, with δ^* proportion of misinformation removed at $t = 0$, before it is observed by any of the users.

https://www.vanityfair.com/news/2020/12/with-the-election-over-facebook-gets-back-to-spreading-misinfor and also https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

[18] Alternately, it can "tag" the article in question as disputed by outside sources, with analogous implications. See, for example, Facebook's policies leading up to the 2020 election on labeling suspected misinformation: https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/.

[19] We think of δ as being a technology parameter related to how effective the regulator is in identifying misinformation. The assumption that the regulator may make type-I errors but not type-II errors (truthful articles are never misidentified, but misinformation is identified with some probability less than one) is adopted for simplicity.

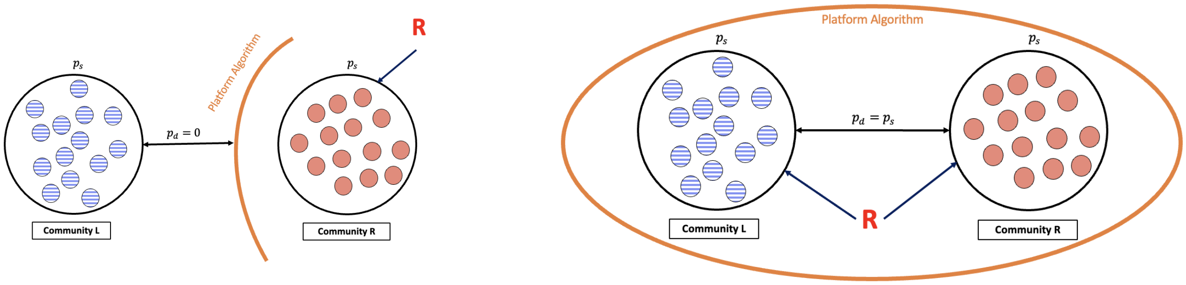
Proposition 2.1.4. *There exist* $0 < \delta_1 < \delta_2 < \delta_3 < 1$ *such that:*

(a) If $\delta \in (0, \delta_1) \cup (\delta_3, 1)$, *then* $\delta^* = \delta$ *is the most effective policy;*

(b) If $\delta \in (\delta_1, \delta_2)$, *the most effective policy sets* $\delta^* < \delta$.

To understand this result, note that censorship has a two-pronged effect. On the one hand, it removes misinformation from circulation and prevents its potential to spread on the platform. On the other hand, it generates an "implied truth" effect for uncensored articles (as empirically identified in Pennycook et al. (2020a)). Bayesian users believe, correctly, that articles are more likely to be truthful when there is censorship of misinformation. In this case, the platform might naturally expand its recommendation filter bubble to generate more engagement, increasing the virality of any remaining misinformation. In some cases, this latter effect may more than offset any gains from the detection and elimination of misinformation.

In part (a), this implied truth effect is not sufficiently powerful, and as a result, both limited (small δ) and highly effective (large δ) censorship lead to better outcomes. Consequently, the policymaker should always censor as much as technologically feasible. In the small δ regime, the sharing network chosen by the platform remains constant and the censorship helps filter out a fraction of the misinformation. In the large δ regime, censorship can remove most of the misinformation, which is the most effective policy in any sharing network, including the one selected by the platform. In the intermediate censorship regime, however, more censorship might exacerbate the spread of misinformation. As we illustrate in the following example, intermediate censorship may create such a serious backlash that it exacerbates the spread of misinformation relative to no censorship.



(a) Low censorship.

(b) High censorship.

Figure 2-3. Optimal Platform Sharing Networks for Example 2.1.1 under Censorship Policies.

Example 2.1.1. Let us consider the two-island setup depicted in Figure 2-2 from Section 2.1.4, where there are $N/2$ left-wing agents with belief $b_L = 5/12$ and $N/2$ right-wing agents with belief $b_R = 7/12$. Let us consider an article with a reliability score indicating it is equally likely to contain misinformation or to be truthful ($\phi(r) = 1/2$), but which is perfectly informative about the state θ ($p = 1$ and $q = 0$).

We assume that $u = c = 1$ and $\kappa = 1/(2N)$ so the payoff from sharing for agent i is given by $U_i = (2\pi_i - 1) + (S_i - D_i)/(2N)$, where π_i is agent i's posterior belief that the article is truthful conditional on reliability and message $m = R$. At the same time, we assume $\tilde{u} = 1$ and $\tilde{c} = 0$, so the payoff from disliking is $1 - \pi_i$ (which by nature of $\pi_i \geq 0$ is always a better response than ignoring).

With no censorship policy ($\delta = 0$), the optimal platform sharing network is given by Figure 2-3a, where the algorithm applies a filter bubble to the right-wing island, shielding the left-wing island from receiving the content. The article spreads among $N/2$ proportion of the population. Once a censorship policy is adopted, the implied truth effect will replace $\phi(r)$ with $\tilde{\phi}(r) = \frac{\phi(r)}{\phi(r)+(1-\delta)(1-\phi(r))}$, leading to a higher value for π_i on both the left and right-wing islands. Consider three separate regimes:

1. *Limited censorship*: With limited censorship ($\delta < 2/7$), the optimal platform sharing network remains the same as in Figure 2-3a. However, the virality of misinformation declines to $(1 - \delta)N/2 < N/2$, and thus overall misinformation is reduced.

2. *Intermediate censorship*: With a more aggressive censorship policy ($2/7 < \delta < 1/2$), the optimal platform sharing network switches to the one shown in Figure 2-3b (maximal connectivity), but still does not filter out most of the misinformation. The platform selects a more expansive sharing network because, with censorship, platform users correctly believe that any given content is less likely to contain misinformation, even counter-attitudinal messages. The platform responds to this by moving from a maximally homophilic network to one with maximal connectivity (as per Theorem 2.1.3). The resulting virality of misinformation then becomes $(1 - \delta)N$, which is greater than $N/2$ for all $2/7 < \delta < 1/2$. In this range, censorship is worse than no censorship policy at all.

3. *Highly effective censorship*: With a censorship policy that can accurately detect most misinformation ($1/2 < \delta < 1$), the policy reduces misinformation, even though the

platform again adjusts its algorithms in response to censorship in order to increase the virality of undetected misinformation. ∎

Provenance

Next, we consider a policy that requires the platform to reveal the original context or *provenance* of a piece of content. For example, provenance may point the user to a peer-reviewed medical study or the full discourse from which a quote was pulled. Such a policy allows users to verify (or "fact-check") social media content easily and quickly.

We model a provenance policy by allowing users to fact-check the article before making their share and dislike decisions. We assume that revealing provenance allows each agent to identify misinformation with (across-user independent) probability $\rho \in (0,1)$; a truthful article is never misidentified as misinformation. Hence, more effective provenance policies allow a greater fraction of users to quickly identify misinformation.[20]

Proposition 2.1.5. *There exist* $0 < \rho_1 < \rho_2 < \rho_3 \leq \delta_3 < 1$ *such that:*

(a) If $\rho \in (0, \rho_1) \cup (\rho_3, 1)$*, then* $\rho^* = \rho$ *is the most effective policy;*

(b) If $\rho \in (\rho_1, \rho_2)$*, then* $\rho^* < \rho$ *is the most effective policy.*

Moreover, a provenance policy with $\rho \in (\delta_3, 1)$ *is more effective than a censorship policy with* $\delta = \rho$*.*

The result is similar to Proposition 2.1.4: soft and strong provenance policies are always effective, but moderate provenance policies can exacerbate the spread of misinformation.[21] The proposition also establishes that provenance policies are in some sense more effective than censorship policies when implemented well. Decentralized fact-checking reduces the likelihood of type-I errors (misidentifying misinformation as truthful) that can result in large share cascades similar to those in Example 2.1.1. Because multiple users are independently

[20] In practice, certain demographic groups, such as users over 65 years old, appear more likely to accept (blatant) misinformation, perhaps because of poor media interpretation skills (see Grinberg et al. (2019) and Guess et al. (2019)). Thus, provenance policies which provide a less clear pathway to fact-checking may lead certain social media users to make type-I errors.

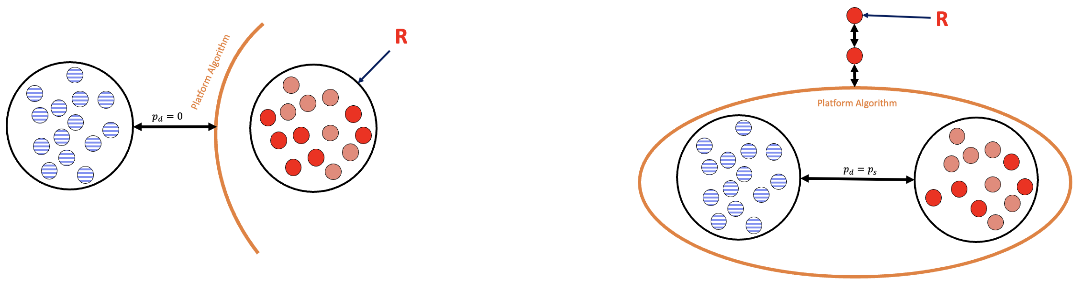
[21] Soft provenance policies are closely related to accuracy nudging interventions, where users are prompted to think carefully about the accuracy of content before sharing. These have been empirically shown to have positive impacts on reducing the spread of misinformation, see Pennycook et al. (2021b) and Pennycook et al. (2020b).

assessing veracity through the provenance channel, misinformation will tend to be stopped as it is checked along various paths in the sharing network. Strategic complementarities further amplify this effect: because users are aware that the provenance policy may allow others downstream to identify misinformation, they are also more cautious themselves in sharing low-reliability content. That being said, provenance policies are not always superior to censorship policies, as illustrated by the following example.

Example 2.1.2. Let us consider the setting of Example 2.1.1, with the slight amendment that the $N/2$ right-wing agents are split into $N/4$ extreme right-wing agents (with belief $b_{RR} = 3/4$) and $N/4$ moderate right-wing agents (with belief $b_R = 7/12$). It is straightforward to verify that the profit-maximizing sharing network for the platform with no policy is still the same as in Example 2.1.1 (Figure 2-4a), and that a censorship policy of $\delta = 3/16$ has the same effect as before (in particular, it is more effective than no policy because $\delta < 2/7$).

Let us now consider a provenance policy with $\rho = 3/16$ (which in this case is in the range (ρ_1, ρ_2)). Here, the following sharing network increases engagement relative to the network in Figure 2-4a: agent 1 is connected to agent 2, who is connected to a clique of the other $N - 2$ agents, as shown in Figure 2-4b.[22] For all agents $i \in \{3, \ldots, N\}$, conditional on the article reaching them, their belief $\tilde{\phi}(r)$ about the article's veracity is greater than under censorship (with $\delta = 3/16$), since two independent fact-checks with $\rho = 3/16$ each have not detected it as misinformation. As a result, all agents in the clique of Figure 2-4b will blindly share the article, correctly assuming that it has likely been already fact-checked. Expected user engagement with misinformation in this case is $(1-\rho) + (1-\rho)^2 + (1-\rho)^3(N-2) > N/2$, for $\rho = 3/16$.

[22] Notice that this sharing network is not in the class of island networks we have focused on so far. In terms of Proposition 2.1.5, when $\rho \in (0, \rho_1) \cup (\rho_3, 1)$, the platform's choices always lie within the class of island networks, but not necessarily when we are outside of this range.



(a) No provenance.

(b) Intermediate provenance.

Figure 2-4. Optimal Platform Sharing Networks for Example 2.1.2 under Provenance Policies.

Consequently, the provenance policy with $\rho = 3/16$ is worse than no policy at all, which is in turn worse than a censorship policy with $\delta = 3/16$. ∎

This example illuminates the potential weakness of provenance policies: when imperfectly implemented, they may be less robust than censorship. Because users presume others before them have fact-checked, they do not make independent judgments based on the reliability of the content. This observation is related to the literature on informational cascades and herding (for example, see Banerjee (1992), Bikhchandani et al. (2021), and Chen and Papanastasiou (2021)). When provenance policies are enacted, agents may excessively follow the sharing decisions of those before them instead of making independent inferences about content veracity before sharing. This "herding" opens the door for share cascades that may increase the virality of misinformation.

Performance Targets

Another possible regulation, which has recently been proposed by social media platforms (see Bickert (2020)), is to set *performance targets* that limit the amount of misinformation. However, since the monitoring and removal of misinformation is imperfect, a performance target allows the platform some leeway to have "bad" content on their site while still requiring a degree of accountability.

While in Sections 2.1.6 and 2.1.6 we considered policies where, respectively, the regulator and the users were responsible for removing content, a performance target transfers the burden of content removal to the platform itself. We assume that the platform can discard content (by not recommending it to any user), and in doing so, forfeits any potential engagement this content may have with the users.

We assume the regulator sets a performance target of λ, which requires the proportion of misinformation shares (to total shares) on the platform to fall below λ.[23] The regulator enforces this performance target by auditing the platform and sampling the content to verify that it meets the required standard. Formally, we assume that the regulator has an auditing

[23] This metric for performance comes from Facebook's own statements on platform standards: "Regulators could say that internet platforms must publish annual data on the 'prevalence' of content that violates their policies, and that companies must make reasonable efforts to ensure that the prevalence of violating content remains below some standard threshold" (from Bickert (2020)), with the definition of prevalence being: "We care most about how often content that violates our standards is actually seen relative to the total amount of times any content is seen on Facebook" (from `https://about.fb.com/news/2019/05/measuring-prevalence/`).

technology $\alpha \in (0,1)$, which represents the probability of detecting that the platform has violated its performance target, and if detected, the platform incurs a cost C due to regulatory fines. Moreover, we assume $\phi(r) < \max_{i^*} \mathbb{E}[\mathbf{S}_{i^*}]$, where the virality is with respect to no policy, otherwise the platform may be happy to comply with a performance target that removes all misinformation.

Our next result establishes how stricter performance targets affect the spread of misinformation:

Proposition 2.1.6. *There exists a performance target $\lambda^* \in (0,1)$ such that:*

(a) If $\lambda > \lambda^*$, a stricter performance target (lower λ) is more effective;

(b) If $\lambda < \lambda^*$, a stricter performance target (lower λ) is less effective than λ^*.

This result establishes that when performance targets are lax, making them stricter (reducing λ) always curbs the spread of misinformation. In this region, when held more accountable, the platform removes some of the misinformation in circulation, foregoing the engagement that these contents would have generated. As a result, lower targets therefore align regulator and platform incentives to remove less reliable content.

However, with stricter targets, the incentives of the regulator and platform diverge. In particular, for targets stricter than λ^*, the platform needs to remove more and more content, with an increasingly larger sacrifice in engagement. In this case, the platform may prefer to violate the performance target and this implies that the tightening of the performance target actually backfires.

This analysis also implies that stricter performance targets need to be combined with better auditing or higher penalties for violation. This simple observation goes against the view that harsher punishments should be imposed when the platform fails to meet low targets (because there would be little excuse for violating them), and weaker punishments may be called for with stricter targets (because the platform may fail to meet them even when it tries). Instead, our analysis clarifies that stricter penalties may be necessary for stricter performance targets in order to prevent its incentives diverging from those of the regulator.

Network Regulations

As we saw in Theorem 2.1.3, when unregulated, the platform chooses the island model of Section 2.1.4 with parameters (p_s, p_d). Here, we consider limits on the ideological homophily induced by the platform's algorithm. Suppose the regulator can choose a homophily standard p^*, based on the ratio between within-island links to across-island links. In other words, this standard would force the platform to choose $p_s/p_d \leq p^*$.

Proposition 2.1.7. *There exists $\gamma < \infty$ such that for any $p^* \geq \gamma$, if the regulator imposes a homophily standard p^*, then (i) the platform chooses the island model with $p_s/p_d \leq p^*$; and (ii) the virality of misinformation is reduced.*

The regulator can thus reduce misinformation by imposing a homophily standard on the sharing network of the platform. This standard prevents the type of extreme homophily we saw in Theorem 2.1.3(a) and forces the platform to choose an algorithm that shares content across ideological groups. This policy is related to the "ideological segregation standard" proposed in Sunstein (2018), which aims to restrict the extent to which content is curated specifically to the ideology or interests of a specific group of users. Such standards ensure that echo chambers are broken and users of differing ideology interact more frequently, limiting the spread of misinformation.

We finally note that the regulation in Proposition 2.1.7 is not always binding for the platform. As Theorem 2.1.3(b) demonstrated, with highly reliable content, the platform maximizes engagement by implementing a maximally-connected sharing network, so the homophily constraint is moot. However, when the article is less reliable, the regulation will bind and the platform will be forced to maintain a minimum level of connectivity between different subgroups.

2.1.7 Conclusion

This paper has developed a simple model of the spread of misinformation over social media platforms. A group of Bayesian agents with heterogeneous priors receive and share news items (articles) according to a stochastic sharing network, determined by the social media platform. Articles may be truthful and informative about an underlying state, or may contain misinformation, making them (weakly) anti-correlated with the underlying state. Upon receiving an article, an agent can decide to share it with others, ignore it, or actively call out another agent for propagating misinformation ("dislike"). Misinformation spreads when agents share articles expecting positive social media feedback and little negative reactions.

Though simple and parsimonious, the model encapsulates several rich strategic interactions. Agents receive utility from sharing truthful articles and not misinformation, but also enjoy peer engagement with shared content. The ideological congruence between an agent and those in her sharing network, which we capture with the notion of homophily, is critical for sharing decisions. Because individuals are more likely to dissent against articles that disagree with their prior beliefs, an agent will be more cautious in sharing articles that disagree with the views of those in her sharing network.

We provide several comparative static results. Some of those are intuitive, though still useful for interpreting a range of results in the emerging empirical literature on social media and misinformation. For example, we find that while misinformation typically spreads less than truthful content (holding all else constant), more sensational content tends to be shared more. Moreover, when misinformation is correlated with sensationalism, the rapid spread of misinformation can be problematic.

Of particular interest are comparative statics with respect to homophily. We show that when there is a highly-reliable article, an increase in homophily reduces the virality of content. Because this article is unlikely to contain misinformation, it is of broad appeal to a wide range of social media users, independent of ideology. An increase in homophily then reduces the extent to which this article can spread throughout the sharing network. The implications of homophily for low-reliability articles are very different, however: in a well-connected network, such articles will be disliked and stopped by users who disagree with their message, and anticipating this behavior and the loss of reputation they can suffer from spreading misinformation, even those who agree with their message would not share them widely. In contrast, high homophily creates echo chambers, where users share low-reliability messages aligned with their beliefs, because they understand that there are few negative reputational consequences from doing so. Misinformation contained in low-reliability articles can then spread virally in these echo chambers.

Our framework enables a tractable study of platform incentives in designing algorithms that determine who shares with whom. To do this, we assume that the platform aims to maximize user engagement (which is a good approximation to the objectives of major social media platforms such as Facebook or Twitter). Our main result is a striking one. When an article is highly reliable, the platform chooses a sharing network with minimal homophily to maximize the spread and appeal of the content throughout the user community. In contrast to this case, when the relevant articles have lower reliability, the platform chooses a network with maximal homophily and recommends articles to users with aligned beliefs. These articles then spread rapidly in the "filter bubble" the platform's algorithms have created—because now ideologically like-minded individuals know that they are unlikely to be caught sharing misinformation in their extreme echo chambers.

We also study regulations aimed at minimizing the spread of misinformation. Content moderation, for example censoring low-reliability articles, can remove some misinformation. However, it also creates a Bayesian version of "false sense of security" and make agents more confident in the quality of remaining items. Similarly, revealing the provenance of a news item (for example, providing full context for a quote or clearer sources) can be useful, because this additional information allows users to more easily fact-check the content for veracity. However, this intervention can backfire, too, because it generates a type of information cascade: each agent expects others to have fact-checked and becomes more lax in his or her inspection. Performance standards that require platforms to remove a certain fraction of posts with misinformation can also backfire, this time because demanding targets can induce the platform to deviate from the standards, with the hope of not being detected. Finally, we show that regulation of platform algorithms, for example, in the form of ideological segregation standards, can be effective, though need to be well calibrated.

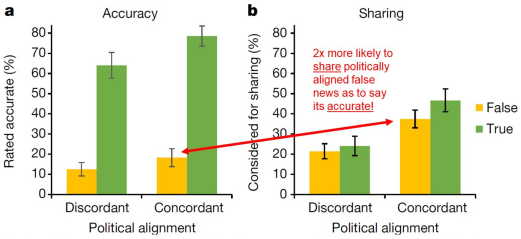
Our framework was purposefully chosen to be simple and several generalizations would be interesting to consider in future work. Most importantly, our assumption that agents are Bayesian rational should be viewed as a useful benchmark. In our setting, it brought out certain new strategic forces—highlighting how social media actions exhibit strategic complementarity and how the degree of homophily alters agents' strategic behavior. Although various behavioral biases and psychological factors appear to be important in social media behavior, we believe that the economic forces we have identified in this paper will continue to apply in the presence of most of these effects, and our Bayesian benchmark enabled us to isolate these forces in a transparent manner. Nevertheless, it remains true that misinformation can be more damaging when agents are boundedly rational, and incorporating such considerations is an important direction for future research. Interesting questions that emerge in this case relate to whether the platform, in addition to designing algorithms that create filter bubbles, may choose strategies that exploit the cognitive limitations of users.

Other theoretical generalizations that might be interesting to consider include extensions to repeated interactions with incomplete information, which would enable agents to also update their beliefs about the ideological position of other agents in their sharing network. Fully endogenizing reputational concerns, taking into account the network position of agents, would be another interesting direction for future research. In this case, the existing reputational capital of an agent will determine how likely she is to risk sharing misinformation. We can also use this extended setup with repeated interactions to study how agents update their initial political views. When there is limited misinformation, agents will gradually learn the true state. In contrast, when there is a significant probability of misinformation, agents will be uncertain about how to interpret articles that disagree with their priors and this may place an upper bound on the speed and possibility of learning (see Acemoglu et al. (2016)).

Despite its simplicity, our model makes several new empirical predictions, most notably related to the non-monotonic effects of homophily and polarization and to platform incentives and algorithmic decisions. Investigating these predictions empirically as well as generating new stylized facts about patterns of these information cascades on social media, is another important area for future research.

2.2 Fighting Fire with Fire: A Model of Misinformation Demand

While social media users are fairly adept at identifying content that contains misinformation, online users still seem to be willing to share co-partisan misinformation more, as shown in Figure 2-5. This observation highlights a disconnect between users' perceptions of truth and



In study 1, $n = 1{,}002$ Americans were presented with a set of 36 headlines (balanced on veracity and partisan slant) and asked to indicate whether they thought the headlines were accurate or whether they would consider sharing them on social media. **a**, The fraction of headlines rated as accurate in the 'accuracy' condition, by the veracity of the headline and political alignment between the headline and the participant. Participants were significantly more likely to rate true headlines as accurate compared to false headlines, whereas the partisan alignment of the headlines had a significantly smaller effect. **b**, The fraction of headlines that participants said they would consider sharing in the 'sharing' condition, by the veracity of the headline and political alignment between the headline and the participant. In contrast to the accuracy condition, the effect of headline veracity was significantly smaller in the sharing condition, whereas the effect of political concordance was significantly larger. Error bars indicate 95% confidence intervals based on robust standard errors clustered on participant and headline.

Figure 2-5. Courtesy of Pennycook et al. (2021b).

their incentives to "share" (i.e., spread) this content to other users.

In Section 2.1, we presented a model of social media users who acted based on social norms and peer encouragement. As we showed in Section 2.1, sharing behavior exhibited strategic complementarity, and so echo chambers among co-partisans led to more aggressive sharing behavior, even for articles containing misinformation. This could explain the divergence between sharing behavior and truth assessment depicted in Figure 2-5.

However, in this section, we present an alternative model that contrasts with the one of Section 2.1. In this model, agents are less interested in peer reactions and care more about influencing others' beliefs to align with their own. As we will demonstrate below, under this formulation, agents are more concerned about the sharing behavior of *counter-partisans* rather than of co-partisans. This occurs because many more voters will have skewed opinions if the opposing ideology is more willing to share misinformation, just to make their point.

In this model, agents may express a demand for pro-attitudinal content, even if it is likely to contain misinformation, because "the ends justify the means" to persuade others to hold similar beliefs. This gives rise to equilibria where users from both ideologies share pro-attitudinal misinformation to combat the sharing of misinformation from the opposing ideology. In these "fight fire with fire" equilibria, content circulating on social media tends more toward misinformation than truthful content, inhibiting the ability of users to *actually* learn the truth from social media content. Policies that lightly nudge users toward considering truth when sharing can destroy these equilibria and push incentives considerably more toward the sharing of only truthful content.

2.2.1 Model

Although considerably different forces arise, the model resembles the one of Section 2.1 closely. There is a true (but unknown) state of the world $\theta \in \{L, R\}$, indicating whether a left-wing or right-wing proposal is better. Two pieces of content are generated, each of which has both a reliability score (i.e., how likely is the headline to contain misinformation) and a message (i.e., what does the headline argue for). This content appears on a social media site consisting of N agents who decide whether to pass this content onto others (i.e., share it or not). Articles are either *truthful* or contain *misinformation*, with truthful articles being more likely to argue for the correct state.

Content. There are two articles A_1 and A_2.[24] Both are generated independently according to the following process:

(i) Reliability: A continuous reliability score (denoted by r_1 and r_2) is drawn on $[0, 1]$ according to some distribution F. This reliability score is perfectly observable.

(ii) Misinformation: The article will either contain misinformation ($\mathcal{M}$) or be truthful ($\mathcal{T}$). The probability that article j is truthful is given by $\phi(r_j)$, where ϕ is an increasing function with $\phi(0) = \underline{z}$ and $\phi(1) = \bar{z}$ with $\underline{z} < \bar{z}$. Whether a given article contains misinformation is not observed.

(iii) Message: If article j is truthful, then the message (denoted m_i) is generated as θ with probability $p > 1/2$ (i.e., it is correlated with the truth). If article j contains misinformation, then the message is generated as θ with probability $q \leq 1/2$ (i.e., it is either orthogonal or anti-correlated with the truth). The message is also perfectly observable.

Agents. Each agent has one of two types $\tau \in \{\mathcal{S}, \mathcal{N}\}$; she is either *strategic* ($\mathcal{S}$) or *naive* ($\mathcal{N}$). Strategic agents are fully rational (Bayesian) agents whereas naive agents behave mechanically, as we describe below. The probability agent i is naive is $\gamma \in (0, 1)$, drawn independently from her prior belief. Agents hold heterogenous prior beliefs about θ. Each agent i holds initial belief $\pi_{i,0}$ about $\theta = R$, which is drawn from distribution H_τ.[25] The game the agent plays involves two phases: the *sharing* phase and the *voting* phase, which occur in that sequence.

Sharing Phase. Time is discrete $t = 0, 1, 2, \ldots$ in the sharing phase. Each article (A_1 and A_2) starts at a strategic agent[26] chosen uniformly at random from the population. When agent i receives an article at time t, she learns from the content presented in the article and then takes a binary sharing action $y_i \in \{\textbf{Share}, \textbf{Ignore}\}$. The action **Share** leads to another agent in the population (chosen uniformly at random) receiving the article at $t + 1$ with probability $1 - \varepsilon$

[24] Our results generalize to any finite number of articles, but our mechanism requires at least two. With multiple articles, agents who receive one article must also consider the strategic sharing behavior of agents who might have received a different article containing different (mis)information.

[25] We allow the distribution of beliefs to depend on τ, e.g., to consider cases where the strategic agents are perhaps more extreme than the naive agents, but our results apply identically if H_τ is independent of τ.

[26] This is for simplicity, but also without loss of generality. Because naive agents act mechanically (and always) elect **Share**, one can begin the analysis at the first strategic agent who receives the article. (This event occurs with high probability whenever ε (defined below) is sufficiently small.)

for some small $\varepsilon > 0$, whereas the action **Ignore** ends the circulation of the article.[27]

We let $\pi_{i,t}$ denote the belief of agent i at time t. Naive agents are inattentive to misinformation, so they always elect **Share** and take all the content seen at face value.[28] In particular, when naive agent i receives an article at time t, she always chooses **Share** and updates her beliefs as:[29]

$$\pi_{i,t+1} \,|\, m_j = R = \frac{p\pi_{i,t}}{p\pi_{i,t} + (1-p)(1-\pi_{i,t})} \tag{2.3}$$

$$\pi_{i,t+1} \,|\, m_j = L = \frac{(1-p)\pi_{i,t}}{(1-p)\pi_{i,t} + p(1-\pi_{i,t})} \tag{2.4}$$

Strategic agents choose y_i and update their belief to $\pi_{i,t+1}$ according to equilibrium play in a sequential equilibrium.

Voting Phase. Note that both articles stop circulating in finite time almost surely (given that $\varepsilon > 0$), so all agents have a terminal belief π_i^* after the sharing phase. After this event, agents "vote" by taking a binary action $a_i \in \{L, R\}$ (e.g., voting for the proposal they think is better). We assume agents get utility 1 for voting for the correct θ and utility 0 otherwise, so agents vote for the state more likely to be true according to π_i^* (i.e., $a_i = L$ if $\pi_i^* < 1/2$ and $a_i = R$ if $\pi_i^* > 1/2$).[30]

Sharing Payoffs. For simplicity, we normalize the payoff of the **Ignore** action to 0. The utility from the **Share** action consists of three components. The first component of sharing utility is that agents receive positive utility for sharing truthful content but have an inherent distaste for sharing misinformation (*sharing for truth*). Formally, the strategic agent receives $B > 0$ if the type of the article is $\mathcal{T}$ and faces a cost $C > 0$ if the type of the article is $\mathcal{M}$.[31] We denote this

[27] This ε probability guarantees that, for a large population, every agent sees at most one article with high probability.

[28] The existence of these naive agents is motivated by empirical work such as Pennycook and Rand (2019). In Section 3.3.4, we consider accuracy nudging as an intervention to make naive agents more aware of misinformation in their sharing decisions and belief updating process.

[29] Note this is the standard Bayesian update for an agent who believes all content is truthful (contains no misinformation) with probability 1.

[30] Observe that agent i's voting payoff is independent of other agents' voting actions or sharing actions (including her own), so one can treat voting actions as a mechanical rule for all (including strategic agents) that solely depend on π_i^*.

[31] In other words, absent of any other sharing incentives, agents get positive utility from sharing truthful content but negative utility from sharing misinformation. This payoff specification is the same as the one assumed in Papanastasiou (2020): the "sharing for truth" component of utility requires a strategic agent to evaluate the likelihood a particular article contains misinformation before sharing it.

(random) truthful sharing payoff as V.

The second component of sharing utility is that the agent wants other agents in the population to vote similarly to her (*sharing for influence*).[32] Let us denote by λ_i as the fraction of votes that agree with agent i (i.e., $\lambda_i = \frac{1}{N}\sum_{k=1}^{\infty} \mathbf{1}_{a_k = a_i}$). We assume that every agent i has a utility function $U(\lambda_i)$, with U monotonically increasing and strictly concave in λ_i.

The final component of sharing utility depends on the catchiness of the article, which we denote by κ. This is an exogenous parameter that captures the desirability of agent i to share the article based solely on how "interesting" the headline is. This can incorporate numerous other social media sharing incentives, such as the desire for re-tweets and/or re-shares from other followers. Agent i's aggregate utility is composed of all three components and given by $V + U(\lambda_i) + \kappa$.

2.2.2 Equilibrium Characterization

Basic assumptions. We make three operating assumptions:

(i) *Large population*: We assume $N \to \infty$. This guarantees that each agent updates her belief at most once, when she receives either article A_1 or A_2 (and does not update at all if she receives neither). Thus, influence amounts to changing $\pi_{i,0}$ to π_i^* conditional on observing an article.

(ii) *No perfectly moderate (strategic) agents*: Agents are either initially left-leaning or right-leaning, and not perfectly moderate. Formally, we assume there exists an open interval $I \equiv (1/2 - \delta, 1/2 + \delta)$ for some $\delta > 0$ where H_s has no support over I (i.e., $\pi_{i,0} \in I$ occurs with probability 0).

(iii) *Anti-correlation lower bound*: The anti-correlation of misinformation is not too large, i.e., there exists $\underline{q}$ (which is a decreasing function in δ) such that $q \in (\underline{q}, 1/2)$.[33] Combined with

[32] This component of the utility function is closely related to the model of Hsu et al. (2020), where agents share content to influence others toward their belief. However, the analysis in Hsu et al. (2020) lacks a full characterization (it largely resorts to limiting cases) and the model has no other sharing incentives (e.g., sharing for truth), which fails to explain many empirical facts, including viral sharing within echo chambers (see Quattrociocchi et al. (2016), Törnberg (2018), and Acemoglu et al. (2022b)).

[33] In practice, this assumption means there are no articles that are so clearly misinformation (and opposing the truth) that, in fact, they argue for the opposite state to a Bayesian agent. For example, a conspiracy theory such as Pizzagate in 2015 (see Fisher et al. (2016)) arguing against Hillary Clinton may have been so ludicrous that it actually bolstered support for Hillary Clinton. Thus, we assume misinformation is mostly just uninformative of the true state θ or mildly anti-correlated. (See Footnote 34.)

assumption (ii), this guarantees that a strategic agent with $\pi_{i,0} < 1/2$ (resp. $\pi_{i,0} > 1/2$) will vote $a_i = L$ (resp. $a_i = R$) conditional on receiving an article advocating for $m_i = L$ (resp. $m_i = R$), regardless of its reliability score.[34]

In what follows, we conjecture the structure of equilibria.

Definition 2.2.1. A cutoff strategy equilibrium is one where there exists a pair of functions $b^*_L(r), b^*_R(r)$ such that for article j:

(i) When $m_j = L$, agents choose **Share** if $\pi_{i,0} < b^*_L(r)$ (and otherwise choose **Ignore**).

(ii) When $m_j = R$, agents choose **Share** if $\pi_{i,0} > b^*_R(r)$ (and otherwise choose **Ignore**).

where $b^*_L(r)$ is increasing in r and $b^*_R(r)$ is decreasing in r.

In other words, Definition 2.2.1 says that for a fixed reliability score, agents are more likely to share content that agrees with their prior than content that disagrees with it. The following result shows all equilibria take this form:

Theorem 2.2.1. *All equilibria are in cutoff strategies and at least one exists.*

Note that Definition 2.2.1 can be also re-written fixing the prior belief $\pi_{i,0}$ and instead employing cutoffs according to the reliability score of the message. In particular, there exists a pair of function $r^*_L(\pi_{i,0}), r^*_R(\pi_{i,0})$ such that:

(i) When $m_j = L$, agents choose **Share** if $r > r^*_L(\pi_{i,0})$ (and otherwise choose **Ignore**).

(ii) When $m_j = R$, agents choose **Share** if $r > r^*_R(\pi_{i,0})$ (and otherwise choose **Ignore**).

Moreover, it can be seen that $r^*_L(\pi_{i,0}) < r^*_R(\pi_{i,0})$ for all $\pi_{i,0} < 1/2$ and $r^*_L(\pi_{i,0}) > r^*_R(\pi_{i,0})$ for all $\pi_{i,0} > 1/2$. Interpreted differently, there is greater demand for low-reliability content when this content happens to agree with the agent's prior beliefs (i.e., pro-attitudinal content) relative to high-reliability content that happens to disagree (i.e., counter-attitudinal content). For a fixed prior $\pi_{i,0}$ the interval $(r^*_L(\pi_{i,0}), r^*_R(\pi_{i,0}))$ (or $(r^*_R(\pi_{i,0}), r^*_L(\pi_{i,0}))$, depending on whether $\pi_{i,0} < b^*$ or $\pi_{i,0} > b^*$) determines the reliability range where ideologically-congruent content, but content possibly containing misinformation, is in higher demand than more reliable content, but which is ideologically opposed.

[34] To see why this anti-correlation lower bound is necessary, note that a low-reliability article that is almost perfectly anti-correlated with the state θ actually argues for not θ, which creates some unnatural incentives. For example, a left-wing agent may share a ludicrous right-wing article to influence other strategic agents that the right-wing proposal is ridiculous, convincing them to vote for L.

Theorem 2.2.2. *There exist extremal equilibria. In other words, for every realization of* (r_1, r_2)*, there are equilibrium cutoffs* $(\bar{b}_L^*, \bar{b}_R^*)$ *and* $(\underline{b}_L^*, \underline{b}_R^*)$ *such that for any other equilibrium cutoffs* (b_L^*, b_R^*)*, we have* $\bar{b}_L^* > b_L^*$*,* $\bar{b}_R^* < b_R^*$*,* $\underline{b}_L^* < b_L^*$*, and* $\underline{b}_R^* > b_R^*$*.*

Under $(\bar{b}_L^*, \bar{b}_R^*)$ there is the maximal sharing equilibrium ("fight fire with fire") and under $(\underline{b}_L^*, \underline{b}_R^*)$ there is the minimal sharing equilibrium ("sharing for truth"). In the maximal sharing equilibrium, there is greater demand for misinformation: for any given message (and reliability), more agents are willing to share the article despite these agents believing there is a greater likelihood of it containing misinformation. Put differently, the maximal sharing equilibrium is the equilibrium where every agent requires a lower reliability to be willing to sharing pro-attitudinal content.

Proposition 2.2.1. *As* $U \to 0$ *almost everywhere, (i.e., only payoff is from sharing truth), then there is a unique equilibrium.*

Multiple equilibria arise due to the strategic incentives to counteract the misinformation that may appear from the other side. Aggressive sharing for influence from one side creates an incentive to more aggressively share from the other side in order to offset the change in beliefs. With sharing for truth only, the best response of each agent does not depend on the actions (or beliefs) of other agents, and thus the equilibrium characterization is straightforward, following similarly from the analysis in Papanastasiou (2020).

2.2.3 Comparative Statics

We consider some comparative statics of interest. In this section, we will focus on an extremal equilibrium, either the "fight fire with fire" equilibrium or the "sharing for truth" equilibrium. Notationally, we denote this initial extremal equilibrium as (b_L^*, b_R^*) and the extremal equilibrium after the shock as $(\tilde{b}_L^*, \tilde{b}_R^*)$ and $(\tilde{b}_L^*, \tilde{b}_R^*)$.

Definition 2.2.2. We say sharing is *monotonically increasing* following a positive shock (for some set of parameters) if $\tilde{b}_L^* > b_L^*$ and $\tilde{b}_R^* < b_R^*$ for all reliability scores (r_1, r_2) of the articles.

Definition 2.2.2 defines an "increase in sharing" as leading to more sharing (and less ignoring) for all types of content. In other words, there is greater demand for misinformation because (strategic) agents are less tethered to high reliability in their sharing decisions. For simplicity in language, we will occasionally say the shock leads to an "increase in sharing" instead of sharing is monotonically increasing.

Throughout, we will make the following assumption about the initial equilibrium cutoffs in the most-sharing equilibrium:

Assumption 2.2.1. For strategic agents, the gap δ around moderate belief $1/2$ (from Assumption (ii) in Section 5.3) is sufficiently large such that any message with maximal reliability does not change the vote of any strategic agent i.

In other words, Assumption 2.2.1 posits that strategic agents are sufficiently opinionated that highly reliable content arguing against their own priors does not move their posterior beliefs of θ enough to change their votes. This guarantees that when strategic agents consider sharing to influence others, they do so to influence naive agents and not other strategic agents.[35] While Assumption 2.2.1 might appear as though it conditions on endogenous equilibrium outcomes, we note that it does not. With a large population, past sharing behavior has no bearing on future sharing behavior (i.e., the process is Markovian), so δ depends only on primitives and not on equilibrium play itself.[36]

Next, we consider comparative statics that hold in all contexts when looking at the minimal and maximal-sharing equilibria. First, we consider both (i) the catchiness of the content and (ii) likelihood the reliability of content and the quantity of misinformation. Because the amount of misinformation is an implicit parameter in our model, we define the following:

Definition 2.2.3. We say that there is an *decrease in misinformation* under $(\tilde{F}, \tilde{\phi})$ relative to (F, ϕ) if $\tilde{F}$ first-order stochastically dominates F and $\tilde{\phi} \leq \phi$ pointwise.

[35] This assumption is made for convenience of the analysis, but the qualitative results do not change if there is potential to influence other (more moderate) strategic agents. The only difference is that the breakdown by ideological sharing and broader sharing (as described below) is no longer as easily determined by comparing the cutoffs to the moderate belief $1/2$.

[36] To see this, observe that for any prior ideological belief $\pi_{i,0}$, the likelihood ψ_i of the content containing misinformation (assessed by agent i), conditional on message m and reliability score r, is always given by:

$$\psi_i(r,m) = \begin{cases} \frac{(q\pi_{i,0}+(1-q)(1-\pi_{i,0}))\phi(r)}{(q\pi_{i,0}+(1-q)(1-\pi_{i,0}))\phi(r)+(p\pi_{i,0}+(1-p)(1-\pi_{i,0}))(1-\phi(r))}, & \text{if } m = R \\ \frac{(q(1-\pi_{i,0})+(1-q)\pi_{i,0})\phi(r)}{(q(1-\pi_{i,0})+(1-q)\pi_{i,0})\phi(r)+(p(1-\pi_{i,0})+(1-p)\pi_{i,0})(1-\phi(r))}, & \text{if } m = L \end{cases}$$

and the posterior belief π_i^* of the state θ is given by:

$$\pi_i^*(\psi_i(r,m),m) = \begin{cases} \frac{((1-\psi_i)p+\psi_i q)\pi_{i,0}}{((1-\psi_i)p+\psi_i q)\pi_{i,0}+((1-\psi_i)(1-p)+\psi_i(1-q))(1-\pi_{i,0})}, & \text{if } m = R \\ \frac{((1-\psi_i)(1-p)+\psi_i(1-q))\pi_{i,0}}{((1-\psi_i)(1-p)+\psi_i(1-q))\pi_{i,0}+((1-\psi_i)p+\psi_i q)(1-\pi_{i,0})} & \text{if } m = L \end{cases}$$

Thus, determining δ such that $(\pi_i^* - 1/2)(\pi_{i,0} - 1/2) \geq 0$ for all $\pi_{i,0}$ in the support of H_s is only a function of primitives. The details of the above calculations are supplied in Appendix **??**.

Note that Definition 2.2.3 states that there is less misinformation present when this content is both more reliable ($\tilde{F} \succeq_{FOSD} F$) and that these reliability scores translate into the same (if not lower) likelihood of containing misinformation. With this definition in hand, we present the following comparative static result:

Proposition 2.2.2. *An increase in catchiness or a decrease in misinformation leads to an increase in sharing.*

The first part of Proposition 2.2.2 establishes that catchy content is more attractive for sharing, despite the fact that this catchy content may be against one's political beliefs (i.e., persuade others to vote contrarily). The second part shows that viral sharing is still more typical in environments where truthful content is more abundant than misinformation. The suggests that while the "fight fire with fire" equilibrium may exhibit a demand for misinformation to offset viral misinformation from the opposing ideology, agents are still disciplined by the quantity of misinformation.

Second, we look at political polarization amongst naive agents in the population. Political polarization has been steadily rising amongst the American electorate (see, for example, Abramowitz (2010) and Pew Research Center (2014)). We consider how the polarization of priors H, using the definition from Acemoglu et al. (2022b), affects sharing in our model:

Definition 2.2.4. We say $\tilde{H}$ is *more polarized* than H if it satisfies the monotone-single crossing property: $\tilde{H}^{-1}(\alpha) - H^{-1}(\alpha)$ is a nondecreasing function in α which crosses (zero) at $\alpha^* = 1/2$ with $H(1/2) = \tilde{H}(1/2) = 1/2$.

Essentially, a polarized society is one where agents become more extreme in their existing prior ideological beliefs. In the case of more polarization within the naive population (i.e., more polarized H_n), we obtain:

Proposition 2.2.3. *An increase in polarization of H_n (naive agents) leads to a decrease in sharing.*

As stated in Proposition 2.2.3, a decrease in polarization amongst naive agents (who are susceptible to being persuaded) decreases the appeal of sharing to influence. As more polarized agents are tethered to their prior belief, sharing provides little influence, as these agents likely would have voted the same way regardless.

For the remaining comparative statics, we categorize the equilibrium (either minimal or maximal sharing) as either a *broad sharing* equilibrium or an *ideological sharing* equilibrium.

A broad sharing equilibrium is one that appeals to the larger community (perhaps because the article has high reliability) in the sense that the sharing cutoff satisfies $b_R^* < 1/2$ and $b_L^* > 1/2$. Conversely, an ideological sharing equilibrium appeals only to one specific ideology, where $b_R^* > 1/2$ and $b_L^* < 1/2$.

For both of these categories, we look at both the *persuasiveness* of the content and the *polarization* of beliefs in society. First, we say that content under $(\tilde{p}, \tilde{q})$ is more persuasive than (p, q) if $\tilde{p} \geq p$ and $\tilde{q} \leq q$. In other words, the article's viewpoint takes a more informative stance on the true state θ. Second, we consider polarization with respect to *strategic* agents (i.e., polarization of H_s).

Broad Sharing. With broad sharing, a large proportion of the population is sharing the content, even if that content is (at least mildly) counter-attitudinal to that agent's ideological belief. In this setting, we obtain the following comparative static:

Proposition 2.2.4. *An increase in persuasiveness or an increase in polarization of H_s leads to a decrease in sharing.*

The most natural way to think about broad sharing is in the case of apolitical content (i.e., cooking videos or dog photos) or with highly reliable content (i.e., from the CDC). This content is very likely to be shared amongst a broad audience. However, once it becomes more politically persuasive (arguing for a certain agenda) or the strategic agents (influencers) become more ingrained in their views, the less likely it is to get spread. Thus, Proposition 2.2.4 establishes that politicization of content or polarization of influencers can turn users off viral content that originally appeals more broadly.

Ideological Sharing. With ideological sharing, where the content is being shared *only* amongst those who are in agreement with the content's message, the comparative statics are reversed, as we establish in the following:

Proposition 2.2.5. *An increase in persuasiveness or an increase in polarization of H_s leads to an increase in sharing.*

To see the intuition for Proposition 2.2.5, consider a politically-charged piece of news (which may or may not contain misinformation) that attracts only a small group of ideologically-congruent users. As its persuasiveness increases, the current cohort of users is more likely to share so that the article can more aggressively persuade naive agents to believe their perspective. Similarly, an increase in the polarization of the influencers has a similar effect: strategic agents more aggressively share to persuade.

Accuracy Nudging for Naive Agents. First, we consider how the fraction of naive agents in the population affects the incentives of strategic agents to share political content:

Proposition 2.2.6. *Sharing is monotonically increasing in the fraction of naive agents in the population γ. Moreover, there exists $0 < \underline{\gamma} < \bar{\gamma} < 1$ such that if $\gamma < \underline{\gamma}$, there is only sharing for truth (i.e., unique equilibrium of Proposition 2.2.1), and if $\gamma > \bar{\gamma}$, there is only sharing for influence (i.e., multiple lattice-ordered equilibria).*

When the fraction of naive agents increases, sharing to influence plays a larger role in the sharing payoff relative to sharing for truth. When agents receive a piece of pro-attitudinal content, they value sharing likely truthful content much less than sharing content that will influence. Because the size of the share cascade is monotonically increasing γ, there is larger sharing for influence payoff when increasing γ (with no effect on sharing for truth).

Next, we consider how accuracy nudging as proposed by Pennycook et al. (2021b) can shift the attention of naive agents toward considering the veracity of the content they see. This has two effects. First, it makes agents less likely to share content that has low reliability. We model this as a *behavioral sharing rule*: the probability that a naive agent shares an article with reliability r is given by an increasing function $\mathcal{R}$. Second, the accuracy nudge may also make naive agents more *cognizant* of the presence of misinformation when updating their belief about θ. In particular, naive agents update their beliefs after seeing article j taking into account the possibility of misinformation:

$$\pi_i^* \,|\, m_j = R = \frac{((1-\psi_i)p + \psi_i q)\pi_{i,0}}{((1-\psi_i)p + \psi_i q)\pi_{i,0} + ((1-\psi_i)(1-p) + \psi_i(1-q))(1-\pi_{i,0})} \tag{2.5}$$

$$\pi_i^* \,|\, m_j = L = \frac{((1-\psi_i)(1-p) + \psi_i(1-q))\pi_{i,0}}{((1-\psi_i)(1-p) + \psi_i(1-q))\pi_{i,0} + ((1-\psi_i)p + \psi_i q)(1-\pi_{i,0})} \tag{2.6}$$

where ψ_i is the likelihood that agent i assigns to the article containing misinformation, given in Footnote 3 (note we have supressed the dependence of ψ_i on r). We assume the accuracy nudge, which affects both sharing and belief updating behavior, is effective on $\beta \geq 0$ proportion of the naive agents.

Proposition 2.2.7. *Sharing (by strategic agents) is monotonically decreasing in the effectiveness of the accuracy nudge β. Moreover, there exists $\beta^* < 1$ such that if $\beta > \beta^*$, then there is a unique "sharing for truth" equilibrium that coincides with the one in Proposition 3.1.1.*

Proposition 2.2.7 provides the interesting insight that while accuracy nudging helps reduce the sharing of unreliable content from naive agents, it also does the same for *strategic* agents. Because naive agents are more critical in their assessment of content accuracy before sharing, supplying future agents with likely misinformation (but which is pro-attitudinal) is less effective. This occurs both because naive agents are less likely to pass the content onto others after them (via the behavioral sharing rule) *and* because unreliable content is less likely to sway their beliefs (by being more cognizant of misinformation). Consequently, strategic agents have relatively lower demand for misinformation and greater demand for truthful content that argues for their perspective.

2.2.4 Experimental Design

Pretest. The design first consists of identifying a large set of headlines that are sufficiently diverse. For this, we ran a pretest that block randomizes questions about a given headline using a few different variants of the same type of question. We are concerned largely with the following attributes of an article: reliability, catchiness, persuasiveness, and message. Here are two such examples of a set of questions for a given headline.

Example 2.2.1. Set of questions per headline #1:

1. Would this headline be more appealing to Republicans or Democrats?

   For all of the following questions, please rate the extent to which you agree or disagree with the following description of the article based on its headline.

2. The article is misleading.

3. The article is credible.

4. You would find this article striking.

5. Other readers would find this article striking.

6. The headline is unconvincing to Democratic voters.

7. The headline is unconvincing to Republican voters.

8. The headline is unconvincing to moderate voters.

Set of questions per headline #2:

1. Would this headline be more favorable to Democrats or Republicans?

   For all of the following questions, please rate the extent to which you agree or disagree with the following description of the article based on its headline.

2. The article is credible.

3. The article is politically biased.

4. You would find this article striking.

5. Other readers would find this article interesting.

6. The headline is convincing to Democratic voters.

7. The headline is convincing to Republican voters.

8. The headline is unconvincing to moderate voters.

Main Experiment Overview. Our main experiment will involve normative blurbs that present statistics from a study on the kind of content users share on social media. After this, users will be prompted to answer various questions about the kind of content *they* would share. In particular, we ask each participant if she identifies as a Democrat, Republican, or Independent and then assign them randomly to one of three treatment groups:

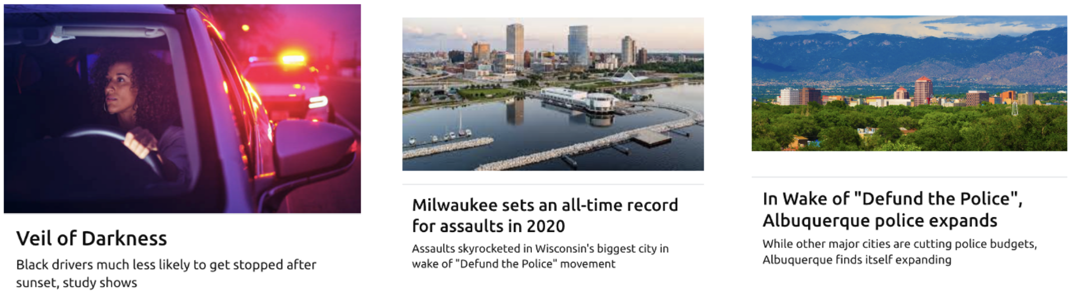
- *In-group treatment*: The participant is presented with a normative blurb that explains a study where pro-partisans share more misinformation than counter-partisans.

- *Out-group treatment*: The participant is presented with a normative blurb that explains a study where counter-partisans share more misinformation than pro-partisans.

- *Control group*: The participant is presented with no blurb prior to the survey on sharing intentions.

The model of Section 2.1 would point in the direction of the in-group treatment leading to more sharing, whereas the model presented in this section would push more in favor of the out-group treatment leading to more sharing.

Defund the Police Survey. There are many empirical and experimental studies that show Republicans share more misinformation more than Democrats (e.g., Guess et al. (2019)), but none that we were aware of that show the opposite. This makes applying the in-group treatment for Democrats (or out-group treatment for Republicans) impossible without deception. To avoid this, we designed a survey that specifically looks at hypothetical headlines associated with the Defund the Police movement, which has much larger Democratic support. We included both a mix of true headlines and false headlines (50% and 50%) as well as a mix of headlines that supported the movement and was opposed (50% and 50%). Examples of these headlines are given in Figures 2-6 and 2-7.



Veil of Darkness

Black drivers much less likely to get stopped after sunset, study shows

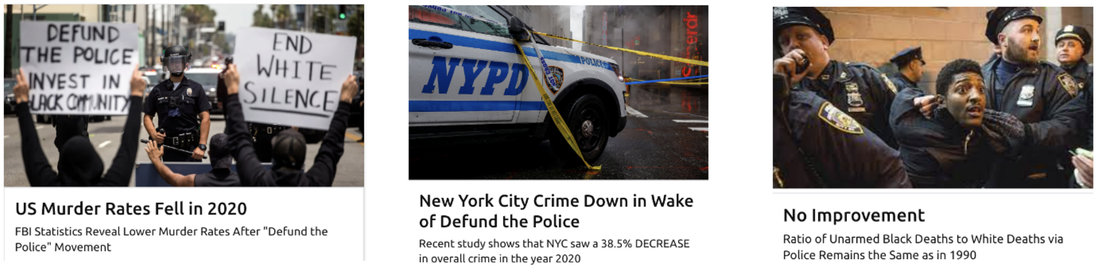
Milwaukee sets an all-time record for assaults in 2020

Assaults skyrocketed in Wisconsin's biggest city in wake of "Defund the Police" movement

In Wake of "Defund the Police", Albuquerque police expands

While other major cities are cutting police budgets, Albuquerque finds itself expanding

Figure 2-6. Sample of truthful headlines.



US Murder Rates Fell in 2020

FBI Statistics Reveal Lower Murder Rates After "Defund the Police" Movement

New York City Crime Down in Wake of Defund the Police

Recent study shows that NYC saw a 38.5% DECREASE in overall crime in the year 2020

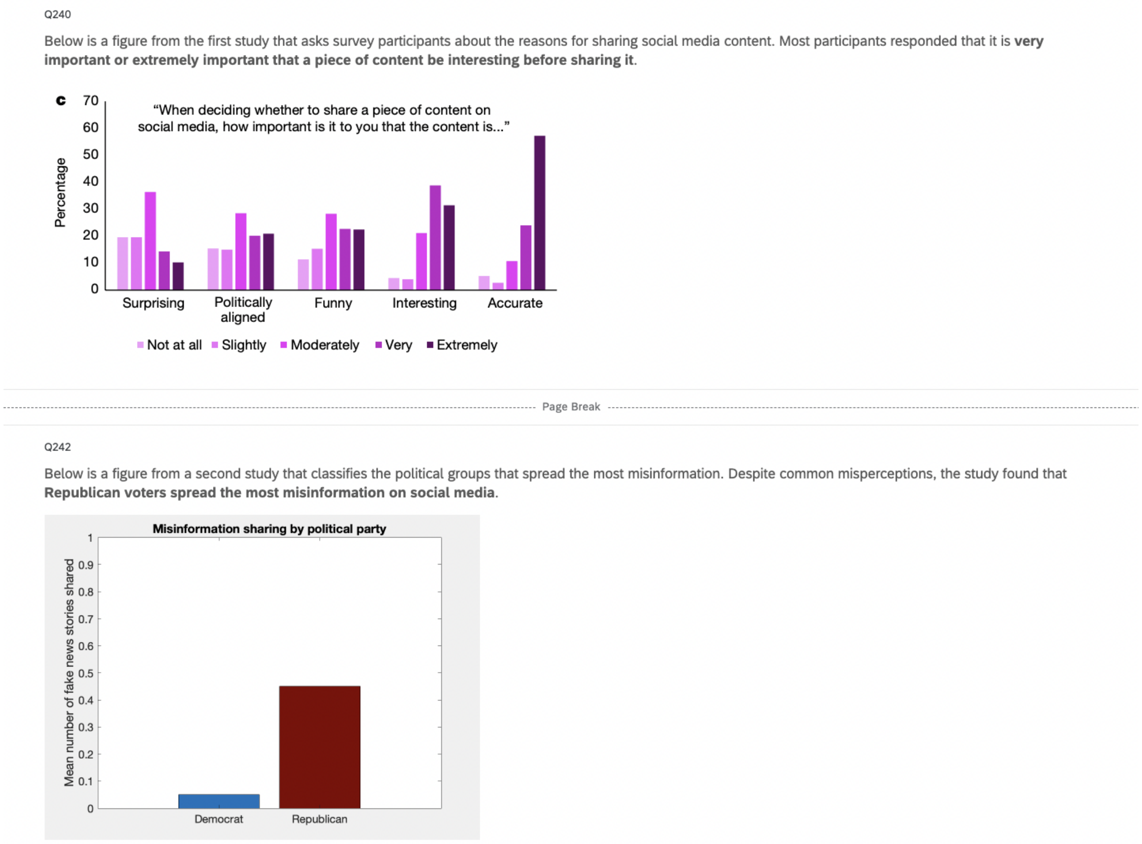
No Improvement

Ratio of Unarmed Black Deaths to White Deaths via Police Remains the Same as in 1990

Figure 2-7. Sample of misinformation headlines.

Main Experiment Flow. After the standard consent and instruction page, each participant will get randomly assigned to one of three normative blurbs:

- *Republicans share more misinformation*: The data from the blurb in Figure 2-8 comes from Pennycook et al. (2021b) and Guess et al. (2019).



Q240

Below is a figure from the first study that asks survey participants about the reasons for sharing social media content. Most participants responded that it is **very important or extremely important that a piece of content be interesting before sharing it**.

Page Break

Q242

Below is a figure from a second study that classifies the political groups that spread the most misinformation. Despite common misperceptions, the study found that **Republican voters spread the most misinformation on social media**.

Misinformation sharing by political party
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean number of fake news stories shared
Democrat
Republican

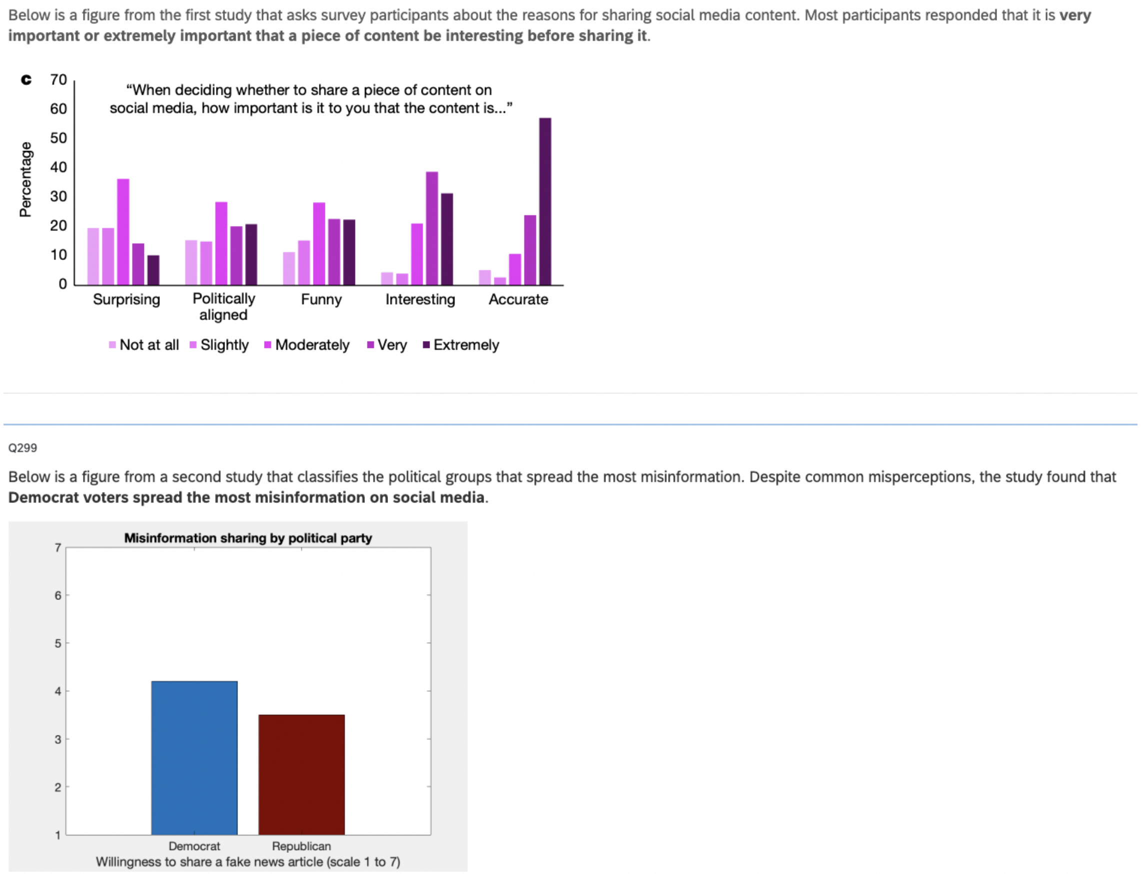
Figure 2-8. Republican normative blurb.

- *Democrats share more misinformation*: The data from the blurb in Figure 2-9 comes from Pennycook et al. (2021b) and the Defund the Policy survey data described earlier.

- *Control group*: The data from the blurb in Figure 2-10 comes from Pennycook et al. (2021b).

To verify participants internalized the blurb shown, we ask them to take the quiz shown in Figure 2-11, which returns to the normative blurb until the participant gets the quiz questions correct. After the quiz, participants will be given a sequence of 10 randomly selected headlines (the same headlines from the pretest) and asked the following questions (with some variations similar to the pretest):

1. How likely are you to share this article on social media?



Below is a figure from the first study that asks survey participants about the reasons for sharing social media content. Most participants responded that it is **very important or extremely important that a piece of content be interesting before sharing it**.

Q299

Below is a figure from a second study that classifies the political groups that spread the most misinformation. Despite common misperceptions, the study found that **Democrat voters spread the most misinformation on social media**.

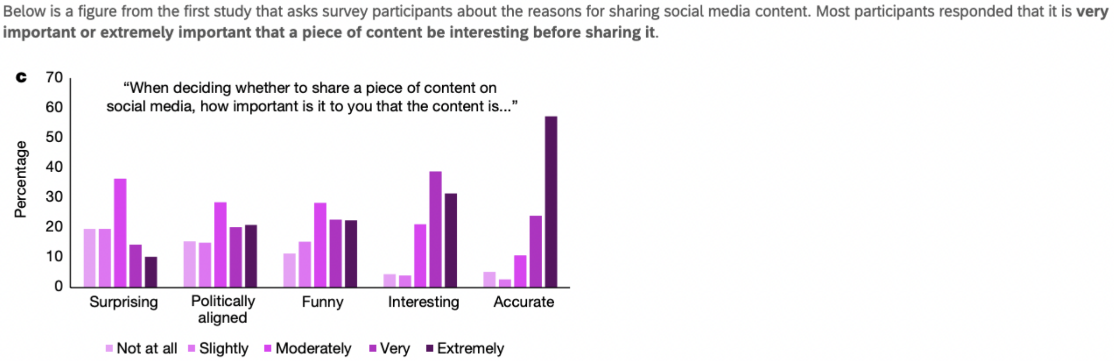
Misinformation sharing by political party
1
2
3
4
5
6
7
Democrat
Republican
Willingness to share a fake news article (scale 1 to 7)

Figure 2-9. Democrat normative blurb.

2. Would you like to read more about this headline?

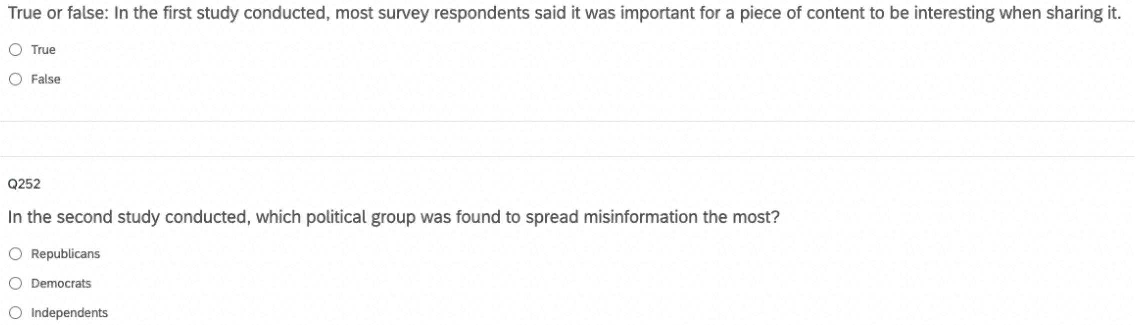
3. How likely are you to like or click on this article?

4. Would you bookmark this article if you saw it online?



Below is a figure from the first study that asks survey participants about the reasons for sharing social media content. Most participants responded that it is **very important or extremely important that a piece of content be interesting before sharing it**.

Figure 2-10. Control normative blurb.



True or false: In the first study conducted, most survey respondents said it was important for a piece of content to be interesting when sharing it.

○ True

○ False

Q252

In the second study conducted, which political group was found to spread misinformation the most?

○ Republicans

○ Democrats

○ Independents

Figure 2-11. Quiz.

Chapter 3

Misinformation: Behavioral Models

We next consider the class of behavioral ("DeGroot") models. While agents' reasoning abilities are assumed to be limited in this setting, this set of tractable models provides additional insights into how misinformation actually shapes social media users' beliefs about incorrect ideas (e.g., that all vaccines are unsafe). Using these models, we can understand the implications for how strategically injected misinformation (often known as "disinformation") can impact society, and specifically how it can negatively affect underprivileged communities with less access to educational resources, are presented in Sections 3.1 and 3.2. Using a reduced-form version of both types of models, I compare the influence of misinformation for both sophistication types of social media agents in Section 3.3, finding that (perhaps surprisingly) sometimes more sophisticated agents can be more susceptible to mislearning.

3.1 Manipulating with Misinformation

In this section, we present a social learning (DeGroot) model based on Mostagir et al. (2022) where agents learn about an underlying state of the world from individual observations as well as from exchanging information with each other. A principal (e.g. a firm or a government) interferes with the learning process by spreading misinformation (or disinformation) in order to manipulate the beliefs of the agents. By utilizing the same forces that give rise to the "wisdom of the crowd" phenomenon, the principal can get the agents to take an action that is not necessarily optimal for them but is in the principal's best interest. We characterize the social norms and network structures that are susceptible to this kind of manipulation, and derive conditions under which a social network is impervious and cannot be manipulated. In the process, we develop a new centrality measure and describe how our model offers insights into designing networks that are resistant to manipulation.

3.1.1 Introduction

People's beliefs can directly impact their actions. These beliefs are usually formed through a combination of individual and social learning, and a large literature details conditions under which learning aggregates beliefs in a way that leads agents to correctly learn an underlying state of the world.

An ability to shape beliefs implies an ability to steer agents towards taking specific actions. In this paper, we consider a social learning environment where agents try to learn an underlying state in order to make a one-time choice between different actions. Agents receive private signals about the state and use these signals in addition to the information they obtain from their neighbors to update their beliefs. A strategic principal is interested in having agents take a certain action, and can try to influence the beliefs in the network by sending costly (and misleading) signals to some of the agents. Agents cannot differentiate whether a signal they are receiving is 'organic' or coming from the principal. Social learning therefore provides a positive externality as agents spread organic news, but also a negative externality as agents unknowingly spread misinformation from the principal. Some agents are *stubborn* – they are endowed with knowledge of the true state and only spread correct information. Being connected to these agents can therefore offer some protection from the influence of the principal.

We say that an agent is manipulated if her beliefs converge to the true state and she takes the correct action in the absence of interference from the principal, but chooses the wrong action due to incorrect beliefs when the principal interferes with the learning process. Our interest is in characterizing the conditions under which the principal can use social learning to his advantage in order to manipulate the agents, and in understanding the social norms and network structures that help or hinder this spread of misinformation in society.

Contribution and Overview of Results. We provide a classification of networks that describes when manipulation is possible. Agents in our model use DeGroot updating to aggregate the beliefs of their neighbors with their own signals in a linear fashion. Theorem 3.1.1 provides a tight characterization of the beliefs of agents under any interference strategy by the principal, and Proposition 3.1.2 proves that these beliefs are related to a novel centrality measure that we call DeGroot centrality. We then show that depending on how agents weigh their own signals, a substantial fraction of the population can be tricked into believing that the underlying state is different from the actual state. Theorem 3.1.3 shows that under mild conditions, extreme societies that are inclined towards herding (agents discount their own signals and put their faith in what other agents think) *or* towards individuality and narcissism (agents discount everything except their own signals) are basically impossible to manipulate. On the other hand, a moderate society whose members use their own beliefs as well as other agents' opinions is the society that is most susceptible to this kind of manipulation.

For these moderate societies, the stubborn agents can help spread the truth about the underlying state, but their ability to do so is limited by the network structure. We classify networks into dense and sparse topologies, and show in Theorem 3.1.4 that dense networks are highly resistant to manipulation: even as the size of the network grows, the presence of a *constant* number of stubborn agents *anywhere* in the network is enough to guarantee imperviousness. By contrast, sparse networks are more susceptible to manipulation, and both the number *and* location of stubborn agents are important for the network to be impervious. In particular, the number of stubborn agents required may grow with the size of the network. If there are not enough stubborn agents, or if there is a sufficient number that are not well-located, then the principal can manipulate almost the entire population by targeting only a fraction of the agents, i.e., it becomes cheaper and easier for the principal to manipulate.

We use the above results to provide a characterization of manipulation in networks that can be represented as a combination of sparse and dense networks, and show the existence of a *phase transition* in Theorem 3.1.5: as the network gets sufficiently dense, all opportunities for manipulation suddenly vanish. Proposition 3.1.7 shows that agents being skeptical about their news source does not necessarily lead to better learning. We then extend our results on several dimensions in Section 3.1.6 and, for the interested reader, provide a numerical study in the appendix that examines the concepts in the paper on data from the advice network described in Jackson et al. (2012).

Related Literature. The agents in our model use DeGroot learning to update their beliefs. DeGroot learning has been extensively studied in several literatures. For example, Golub and

Jackson (2010) give conditions under which beliefs converge to the true state of the world. Jadbabaie et al. (2012) consider agents that update their own information in a Bayesian fashion and aggregate the information of their neighbors in a DeGroot fashion, and their particular formulation of DeGroot agents is closely related to the one we consider in this paper. Bohren and Hauser (2017) examine learning when agents have a misspecified model of the world.

Our model also includes stubborn agents who hold correct beliefs about the state of the world. Opinion dynamics with stubborn agents have been studied in Acemoglu et al. (2013) and Yildiz et al. (2013) among others. The primary differences between our work and these papers is the presence of a strategic principal, which changes the role that these stubborn agents play. In the cited literature, the presence of stubborn agents leads to divergence of opinions and generally hinders learning about the true state of the world. In contrast, the learning difficulty in our model comes from the strategic principal who tries to manipulate beliefs, and the presence of stubborn agents who know the state is always useful for everyone in the network. Nevertheless, as we discuss, even with the positive contribution that these agents provide to the learning process, manipulation might still be unavoidable.

The proliferation of false news on social networks has been the central focus of some recent work. Candogan and Drakopoulos (2020) and Papanastasiou (2020) examine how (Bayesian) agents exchange information on a social network and show how misinformation can spread in these models and what the platform (over which the agents are communicating) can do about it. The existence of fake news in these models is exogenous, i.e., unlike our model, there is no principal or news provider that strategically injects misinformation into the network, and consequently there is no notion of manipulation. The idea that a principal can use social learning to manipulate agents towards taking a certain action has been studied in the context of replicator dynamics in Mostagir (2010). Unlike this work, we examine richer learning dynamics in an environment where some agents consistently spread correct information while others spread their beliefs without critical reasoning, which as Pennycook and Rand (2019) show in recent experimental work, might be one of the primary mechanisms through which misinformation spreads in social networks.

3.1.2 Model

We consider a directed social network with n agents trying to learn a binary state of the world $y \in \{S, R\}$ over time. Time is continuous and agents learn over a finite horizon, $t \in [0, T)$. At time $t = 0$, the underlying state $y \in \{S, R\}$ is drawn, with $\mathbb{P}(y = S) = q \in (0, 1)$.

Organic News News is generated according to a Poisson process with parameter $\lambda > 0$ for each agent i. We refer to this process as *organic news.* For simplicity, we assume agents digest news at the same times $\tau = 1, 2, \ldots$, which correspond to the arrivals of a single Poisson process, but might correspond to different articles (i.e., different messages) for different agents. For all $\tau \in \{1, 2, \ldots\}$, the organic news for agent i generates a signal $s_{i,\tau} \in \{S, R\}$ according to the distribution:

$$\mathbb{P}\left(s_{i,\tau} = S \middle| y = S\right) = \mathbb{P}\left(s_{i,\tau} = R \middle| y = R\right) = p_i \in [1/2, 1) \tag{3.1}$$

i.e., the signal is correlated with the underlying truth. All organic news' signals for agent i are independent across time and across other agents. The value of p_i indicates the richness of agent i's signal, and can be interpreted as her ability to deduce the true state from the facts presented in the organic news. We allow for the possibility that $p_i = 1/2$, so that agent i faces an identification problem and cannot rely on her organic news alone, but instead must rely on others in order to learn the true state.

Principal In addition to the organic news process, there is a strategic principal who may also generate news of his own. We assume, without loss of generality, that the true state is $y = S$ and the principal wants to convince agents of state R.[1] The principal picks an influence strategy $x_i \in \{0, 1\}$ for each agent i in the network. The principal then generates news of his own, which is always signal R, and the influence strategy indicates which agents receive these signals. If the principal chooses $x_i = 1$ for any agent i, then he (principal) generates news according to an independent Poisson process with intensity λ^* which is received by all agents with $x_i = 1$.[2] The principal incurs an upfront investment cost $\varepsilon > 0$ for each agent with $x_i = 1$.

Once again, for simplicity we assume agents digest news at the same rates, so an agent i with $x_i = 1$ receives organic news at time τ with probability $\lambda/(\lambda + \lambda^*)$ and receives the principal's news with probability $\lambda^*/(\lambda + \lambda^*)$, but is unable to differentiate between the nature

[1] This is without loss because if the underlying state is indeed R, then as we establish in Proposition 3.1.1, agents will learn that state without interference from the principal, and therefore he will elect not to intervene.

[2] To simplify our setup, we do not allow the principal to send S messages, vary his influence strategy, or change the intensity of his messages over time. We explore how this affects our results in Section 3.1.6.

of the news. On the other hand, an agent i with $x_i = 0$ always receives organic news. An organic message always follows the distribution in (3.1), whereas a message from the principal always gives a signal of R, i.e., it is misinformation.

The principal can be one of two types. He can either be a strategic type $\mathcal{S}$ or a truthful type $\mathcal{T}$. If the principal's type is $\omega = \mathcal{T}$, we assume he is committed to implementing $x_i = 0$ for all agents; that is, he does not interfere with the learning process. On the other hand, the $\omega = \mathcal{S}$ type of the principal may play any influence strategy $\mathbf{x} \equiv \{x_i\}_{i=1}^n$ over the network.

Agents There are two types of agents in the model. *DeGroot* agents learn about the state by combining both (i) what they read in the news and (ii) what their friends believe about the state. *Stubborn* agents are endowed with knowledge of the true state y at $t = 0$, through being well-educated or knowledgeable about the subject. Stubborn agents will not change their beliefs over time. We denote the set of DeGroots as D and the set of knowledgeable stubborn agents as K. The total population in society is denoted by n, with m denoting the number of stubborn agents in that society. Unlike stubborn agents, DeGroot agents start with prior q about the state y at $t = 0$, and must use their own signals combined with social learning to try and learn the state. Specifically, every DeGroot agent:

(a) uses a simple learning heuristic to update beliefs about the underlying state from other agents.

(b) believes all signals arrive according to a Poisson process with intensity λ and all signals are independent over time with $\mathbb{P}\left(s_{i,\tau} = y\right) = p_i$ (i.e., takes the news at face value).

The implicit assumption in the DeGroot learning process is that DeGroots are not aware of a principal who might be tampering with this process and sending misinformation. DeGroots absorb *all* news as if it is coming from organic sources. We relax this in Section 3.1.6, where agents try to simultaneously learn how trustworthy their news sources are, and can appropriately discount their own news if they suspect it is interfered with.

Formally, let $\pi_{i,t} \in \Delta(\{S, R\})$ represent the belief of agent i about the underlying state at time t. Given history $h_{i,t} = (s_{i,1}, s_{i,2}, \ldots, s_{i,\tau^*})$ up until time t (where τ^* is the last message received before t), each agent forms a personal belief about the state according to Bayes' rule. Let $z_{i,t}^S$ and $z_{i,t}^R$ denote the number of S and R signals, respectively, that agent i received by time

t (where $z_{i,t}^S + z_{i,t}^R = \tau^*$); then DeGroot agent i has direct "personal experience":

$$\text{BU}(S|h_{i,t}) = \frac{p_i^{z_{i,t}^S}(1-p_i)^{z_{i,t}^R}q}{p_i^{z_{i,t}^S}(1-p_i)^{z_{i,t}^R}q + p_i^{z_{i,t}^R}(1-p_i)^{z_{i,t}^S}(1-q)}$$

and $\text{BU}(R|h_{i,t}) = 1 - \text{BU}(S|h_{i,t})$. The experience function BU represents the direct contribution of the observed signals into agent i's belief, and is related to the personal Bayesian update in Jadbabaie et al. (2012). It is the belief any fully Bayesian agent would hold about the state y *in isolation* and with no knowledge of principal interference.

DeGroot agents also form beliefs by talking to (and exchanging beliefs with) their neighbors after every unit of time. For all agents i, there are weights θ_i, α_{ij} such that agent i holds belief $\pi_{i,t}$ for all $t \in \{1, 2, \ldots\}$:

$$\pi_{i,t+1} = \theta_i \text{BU}(h_{i,t+1}) + \sum_{j=1}^{n} \alpha_{ij}\pi_{j,t}$$

where $\theta_i + \sum_{j=1}^n \alpha_{ij} = 1$. As convention, we assume the link $i \to j$ indicates that agent j is a neighbor of i (i.e. i listens to j) but not necessarily vice versa. We refer to this as the *DeGroot update* (DU) process.

Network Structure Each agent i has a neighborhood $N(i) \subset \{1, \ldots, n\}$ that consists of other agents she listens to in every period (i.e., her "friends"). Note that because stubborn agents do not change their beliefs over time, their neighborhoods are immaterial. On the other hand, each DeGroot agent i's neighborhood is specified by her weights $(\theta_i, \{\alpha_{ij}\}_{j=1}^n)$, with larger weights representing stronger connections. In matrix notation, we can represent the *influence* of the social network as:

$$\mathbf{W} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{A}_{DK} & \mathbf{A}_{DD} \end{pmatrix}$$

where $\mathbf{A}_{DK}$ is the DeGroot by stubborn agent weight matrix, given by entries $\{\alpha_{ij}\}_{i\in D,\, j\in K}$, and $\mathbf{A}_{DD}$ is the DeGroot by DeGroot agent weight matrix, given by entries $\{\alpha_{ij}\}_{i,j\in D}$. We refer to this social network as G, and denote the weights w_{ij} from matrix $\mathbf{W}$.

It is also sometimes insightful to look at the unweighted representation of G, which we denote by social network $\mathbf{G}^*$. The unweighted social network is a binary relation between pairs of agents representing whether agent i listens to agent j *at all.* By convention, a link $i \to j$ exists in the undirected social network $\mathbf{G}^*$ if and only if j is in i's neighborhood, i.e., $j \in N(i)$, because $\alpha_{ij} > 0$.

Payoffs At time $t = T$, each agent chooses an action $a_i \in \{S, R\}$. Payoffs for the strategic principal and agent are given in Table 3.1.[3] The first entry in a cell is the principal's payoff while the second is the agent's payoff (so for example, the top-left cell corresponds to the case when the state is R and the agent chooses action R. This gives the principal a payoff of 1 and the agent a payoff of $(1 + b)$).

We assume that $b \in (-1, 1)$ so that agent i would match her action a_i with the state y if she knows the state with certainty. Otherwise, the parameter b captures any asymmetry in the payoffs between the two states.[4] Recall that, on the other hand, the principal always prefers that agents take action R instead of action S, and so has an incentive to convince agents of $y = R$ (when the state is in fact $y = S$). Let $u_i(y, a_i)$ denote the payoff of agent i when the state is y and she takes action a_i; $u_i^p(a_i)$ is the payoff for the principal at agent i (which only depends on that agent's action). The total payoff for the principal is given by $u^p(\mathbf{a}) = \sum_{i=1}^n u_i^p(a_i)$, which is the summation of the payoffs from period-T actions of all n agents (where $\mathbf{a} \equiv \{a_i\}_{i=1}^n$). We denote by $c(\mathbf{x}) = \sum_{i=1}^n \varepsilon x_i$ the cost of the principal for implementing the network influence strategy $\mathbf{x}$ at $t = 0$. The principal has total payoff given by the difference between her future utility (via the actions of the agents) and the cost of the network influence, $u^p(\mathbf{a}) - c(\mathbf{x})$. Agents simply choose an action that maximizes their utility $u_i(y, a_i)$ given belief of the state $\pi_{i,T}$.

Note that the action of the stubborn agent is always S and yields her a payoff of 1. On the other hand, a DeGroot agent will take action S if and only if her terminal belief about state S, $\pi_{i,T}(S)$, exceeds the threshold $(1 + b)/2$, because then action S gives her more (expected) utility than action R. The principal chooses his optimal influence strategy, denoted $\mathbf{x}^*$, based on the expectation of his utility from the actions a that the agents take at time T. Observe that the optimal influence strategy $\mathbf{x}^*$ may or may not be unique.

[3] An example to help visualize this payoff table is the following: the states of nature S and R can be mapped to whether a particular vaccine is safe (state S) or risky (state R). Similarly, an agent's actions can be thought of as analogous to "vaccinate" (action S) and "not vaccinate" (action R). In this sense, a player wants to match her action to the state, e.g. taking action S when the state is S indicates vaccinating when the vaccine is safe.

[4] For instance, it may be more costly to vaccinate your child if vaccines do have adverse effects than it is to not vaccinate even if they are safe.

		Agent	
		R	**S**
State y	**R**	$1, 1+b$	$0, 0$
	S	$1, b$	$0, 1$

Table 3.1. Terminal payoffs when the strategic principal wants agents to take action R. The parameter b is in $(-1, 1)$.

3.1.3 Learning Dynamics and Centrality

In this section, we characterize the learning dynamics and terminal beliefs of agents in the presence of the principal's interference. Our key insight is the relationship between the limiting beliefs of the agents and a novel centrality measure that we call *DeGroot centrality*. This measure captures an agent's susceptibility to misinformation by computing her centrality amongst other (DeGroot) agents who update their beliefs using possibly misinformative signals sent by the principal.

Learning

We aim to understand the asymptotic learning dynamics that emerge for a given (arbitrary) network strategy $\mathbf{x}^*$ of the principal. We provide a closed-form expression for DeGroot terminal beliefs as a function of the chosen influence vector $\mathbf{x}^*$. These terminal beliefs induce actions for each DeGroot agent i at T, which in turn provides an expression for whether agent i mislearns the state under $\mathbf{x}^*$.

When the network consists entirely of stubborn agents, the principal is unable to get anyone to take the incorrect action. On the other hand, when the network consists of all DeGroot agents, generally it will be possible to convince DeGroot agents of the wrong state, as long as the influence cost ε is not too large. The interesting case happens in the mixed environment where both DeGroot and stubborn agents co-exist. In this setting, there are two opposing forces: (i) the stubborn agents who know and can communicate the correct state information, and (ii) the DeGroot agents who may confound the learning process through simple learning heuristics. Our interest in in whether the principal can effectively utilize the second force to his benefit, despite the presence of the first.

We make some assumptions about the rate of information arrival, the informativeness of organic signals, and the network structure:

Assumption 3.1.1. Each of the following hold:

(i) *Amenability to mislearning*: For all agents i, $p_i < \frac{\lambda^*+\lambda}{2\lambda}$.

(ii) *Strong connectedness*: For every two agents i, j in unweighted social network $\mathbf{G}^*$, there exists both a directed path from i to j and from j to i.

(iii) *Irrelevance of noise*: For every DeGroot agent i, θ_i is positive if and only if $p_i > 1/2$.

(iv) *Identifiability*: There exists some agent i where $p_i > 1/2$ (i.e., state R and state S can be identified by at least one agent from solely organic news).

The first part of the assumption ensures that if agents are left in isolation, and the principal attempts to corrupt their signals, then it is impossible for agent i to uncover the truth simply from performing Bayesian updating on her own signals (and ignoring others). However, if the agent utilizes social learning, she may be able to learn the true state. The second part of the assumption requires that the beliefs of any one agent can reach (or influence) any other agent, albeit indirectly through others. The third part requires that all agents in the network listen to the news they receive if and only if their organic news is believed to be meaningful. Lastly, we assume the organic news contains valuable information for at least one agent, otherwise learning would be impossible with all DeGroot agents, even without principal interference.

To understand the role of the network structure in the principal's problem, we need to characterize asymptotic learning for the DeGroot agents. Let y' denote an arbitrary state. We write $h_{i,t}(\mathbf{x}^*)$ as the (random) history of news (both organic and inorganic) up until time t induced by the principal's action $\mathbf{x}^*$ (which, naturally, depends on his type). We first establish that the personal experience component of all agents converges almost surely for a long learning horizon:

Lemma 3.1.1. *The personal-experience Bayesian update term BU$(S|h_{i,t})$ converges almost surely to a constant BU$(S|h_{i,\infty}(\mathbf{x}^*)) \in \{0, q, 1\}$ as $T \to \infty$.*

Given Lemma 3.1.1, we observe that in the limiting case (i.e., large t), the beliefs of the DeGroot agents approximately follow:

$$\boldsymbol{\pi}_{t+1}(y') = \mathrm{BU}(\mathbf{h}_\infty(\mathbf{x}^*))(y') \odot \boldsymbol{\theta} + \mathbf{W}\boldsymbol{\pi}_t(y')$$

where the matrix $\mathbf{W}$ is the influence matrix from Section 3.1.2, $\boldsymbol{\theta} = (\mathbf{1}_K, \theta_{m+1}, \ldots, \theta_n)$, $\odot$ denotes the element-by-element product, and $\mathrm{BU}(\mathbf{h}_\infty(\mathbf{x}^*))(y')$ is the vector of converged personal-experience beliefs of state y', per Lemma 3.1.1 (with the convention that $\mathrm{BU}(S|h_{i,\infty}) = 1$ for all stubborn agents). Given this formulation, we have the following asymptotic result for the beliefs of DeGroot agents:

Theorem 3.1.1. *Under Assumption 3.1.1, the beliefs of the agents about state y' converge almost surely to:*

$$\boldsymbol{\pi}_t(y') \overset{a.s.}{\to} (\mathbf{I} - \mathbf{W})^{-1}(\mathrm{BU}\left(\mathbf{h}_\infty(\mathbf{x}^*)\right)(y') \odot \boldsymbol{\theta})$$

for any principal action $\mathbf{x}^*$.

First, we look to characterize beliefs in the baseline case of a truthful principal (i.e., $\omega = \mathcal{T}$), i.e., $\mathbf{x}^* = \mathbf{0}$. We obtain the following result, which is similar to the findings in Jadbabaie et al. (2012):

Proposition 3.1.1. *If Assumption 3.1.1 holds, then all agents learn the true state almost surely (i.e., $\pi_{i,t}(S) \overset{a.s.}{\to} 1$ for all i) when the principal is truthful.*

Without a strategic principal, learning occurs despite the fact that DeGroot agents are only updating their beliefs using naive learning heuristics. We now turn our attention to whether it is possible for (some) agents to mislearn the state when the strategic principal plays $\mathbf{x}^* \neq \mathbf{0}$. This is the main focus of the remainder of the paper.

Under Assumption 3.1.1, we know by Lemma 3.1.1 that $\mathrm{BU}(\mathbf{h}_\infty(\mathbf{x}^*))(R) \odot \boldsymbol{\theta}$ converges almost surely to the vector:

$$\left(\gamma \equiv \begin{pmatrix} \mathbf{0}_K \\ \mathbf{x}^*_D \end{pmatrix}\right) \odot \boldsymbol{\theta}$$

where the subscripts K and D denote the intervention vector $\mathbf{x}$ associated with those types of agents. In other words, if an agent i is DeGroot and receives misinformation signals from the principal (i.e., $x_i = 1$), then we write $\gamma_i = 1$ and otherwise we write $\gamma_i = 0$. The (limit) beliefs of stubborn agents are naturally a point-mass on the true state (i.e., zero belief on $y' = R$). This allows a succinct representation of the limiting beliefs of the incorrect state $y' = R$ for all agents:

$$\boldsymbol{\pi}_t(\mathbf{x}^*) \to (\mathbf{I} - \mathbf{W})^{-1}(\boldsymbol{\gamma}(\mathbf{x}^*) \odot \boldsymbol{\theta}) \equiv \boldsymbol{\pi}_\infty(\mathbf{x}^*) \tag{3.2}$$

Equation (3.2) provides a closed-form expression for the beliefs of the agents in state R when T is large. We note that this expression depends on the network structure, the a priori knowledge of the agents about the state (i.e., agent types), the personal-experience weights $\boldsymbol{\theta}$, and the network action x of the principal (captured through γ). We discuss each of these factors in more detail below.

1. *Personal-Experience Weights*: Each agent i's personal-experience belief update propagates to the beliefs of other agents in society precisely through her weight θ_i, which factors into the expression $\boldsymbol{\theta} \odot \gamma$. In Section 3.1.4, we show the nuances of how increases in θ_i can either help or hurt the spread of misinformation, as a function of other agents' θ_j for all $j \neq i$.

2. *Network Structure*: Recall from Section 3.1.2 that $\mathbf{W}$ represents the influence matrix. The term $(\mathbf{I} - \mathbf{W})^{-1}_{ij}$ represents the entire accumulation of (direct or indirect) influence j has over i.[5] In Section 3.1.5, we focus on how the topology of the social network G shapes how influence propagates through this term.

3. *Agent Types*: The type of the agent (i.e., replacing a DeGroot agent with a stubborn one) has an impact on both $(\mathbf{I} - \mathbf{W})^{-1}$ and $\boldsymbol{\theta} \odot \gamma$. Through the expression $(\mathbf{I} - \mathbf{W})^{-1}$, one can see that a DeGroot agent may not take the incorrect actrion herself but still spreads some of the misinformation she observes from the beliefs of her friends (or in her news). A stubborn agent on the other hand does not propagate nor succumb to misinformation, which limits the influence the principal can have in the population.

The set of agents taking the incorrect action is entirely determined by the principal's optimal choice of $\mathbf{x}^*$ and the resulting limiting beliefs $\boldsymbol{\pi}_\infty$. For some stylized settings, we can characterize the principal's optimal strategy $\mathbf{x}^*$, as well as the set of agents who will take the incorrect action because of $\mathbf{x}^*$, as a function of these model parameters (see Section 3.1.4 and Section 3.1.5). In Appendix B.1.1, we present the general optimization problem of the principal for arbitrary network parameters, as well as some technical results of interest. For an illustration of how these techniques can be visualized in the context of a real-world social network, we refer the reader to the appendix.

[5] To see this, note the term $(\mathbf{I} - \mathbf{W})^{-1}$ can be written in expanded form as $\sum_{\ell=0}^{\infty} \mathbf{W}^\ell$, where each $\mathbf{W}^\ell$ represents how the beliefs of agent i propagate to agent j who is ℓ hops away in the social network.

Manipulation

A main focus of our paper is characterizing the conditions under which an agent chooses the terminal action that matches the underlying state (and therefore maximizes her ex-post payoff) given her belief at time T. Recall from the previous section that the beliefs of all agents converge almost surely to some limit belief, based on which she takes her terminal action. To this end, we define what it means for agent i to be manipulated:

Definition 3.1.1 (Manipulation)**.** Let $\mathbf{x}^*$ be the optimal network influence strategy for the strategic principal. We say that agent i is *manipulated* under the network influence strategy $\mathbf{x}$ if:

1. Agent i's terminal action a_i *does not* match the underlying state when the principal's type is $\omega = \mathcal{S}$ and $\mathbf{x} = \mathbf{x}^*$, almost surely.

2. Agent i's' terminal action a_i *does* match the underlying state when the principal's type is $\omega = \mathcal{T}$ and $\mathbf{x} = \mathbf{0}$, almost surely.

In other words, manipulation of agent i implies that a strategic principal interferes with the learning process, and this causes the agent to mislearn the true state that she would have correctly learned in the absence of such interference (by Proposition 3.1.1). Agents whose beliefs of state R converge to a value higher than $(1-b)/2$ are necessarily manipulated.

To be able to speak about the extent of manipulation (i.e., how many agents are manipulated) when the principal acts optimally, it is important to consider the entire set of optimal strategies for the principal. Let $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ be two optimal strategies for the principal and let k_1 and k_2 denote the corresponding number of manipulated agents at time T. We say the principal's optimal influence strategies are *manipulation-invariant* if for all $\kappa > 0$, there exists T^* such that for all $T > T^*$, the manipulation at horizon T satisfies:

$$\mathbb{P}_{(k_1,k_2)\sim(\mathbf{x}_1^*,\mathbf{x}_2^*)}\left[k_1 = k_2\right] \geq 1 - \kappa$$

for any two optimal strategies $\mathbf{x}_1^*, \mathbf{x}_2^*$. In other words, manipulation is the same under all optimal strategies for the principal if these strategies are manipulation-invariant. We can then state:

Theorem 3.1.2. *There exists a set $\mathcal{P} \subset \mathbb{R}_+^2 \times (-1,1)$ of measure zero,*[6] *such that for all $(\varepsilon, \lambda, b) \notin \mathcal{P}$, the principal's optimal strategies are manipulation-invariant.*

Manipulation-invariance of the principal's strategies guarantees that, with high probability, our welfare analysis (i.e., the number of agents who mislearn the state) does not depend on the strategy the principal chooses or the realization of the signals or actions during the learning process, as $T \to \infty$. Because of Theorem 3.1.2, we can refer without ambiguity to the "number of manipulated agents" in the principal's optimal strategy. Note that it may be possible that *different* agents are manipulated under different optimal principal strategies but the *total number* of manipulated agents remains unchanged.

DeGroot Centrality

We characterize manipulation in an arbitrary network by developing a centrality measure called DeGroot centrality, which is closely related to the familiar eigenvector, Katz-Bonacich, and PageRank centrality measures from the social learning literature. A definition that allows for a simple visualization of DeGroot centrality is based on weighted walks: fix the weighted social network **G**, its matrix representation **W**, and its unweighted counterpart network **G*** (see Section 3.1.2). A *walk* between agents i and j is any directed path $W = i \to v_1 \to v_2 \to \ldots \to v_k \to j$ such that all links exist in the unweighted social network **G***.[7] The *weight* of a walk W is given by:

$$w_W = \prod_{(v_i \to v_{i+1}) \in W} w_{v_i \to v_{i+1}}$$

We say that a walk W is stubborn-avoiding if none of the agents along the walk are stubborn. Let $\mathcal{W}_{ij}$ be the (countable) set of all stubborn-avoiding walks (of any length) between agents i and j.[8] Recall the vector γ from Section 3.3.3; we will refer to this as the *influence* parameter

[6] The condition that $(\varepsilon, \lambda, b)$ must lie outside $\mathcal{P}$ is often referred to as a *genericity condition*, where we say that $(\varepsilon, \lambda, b)$ are *generic* if they satisfy this property. Its purpose is to eliminate knife-edge cases where specifically chosen parameters may make some agents indifferent between multiple actions, but if perturbed just slightly, the agent prefers a unique action. Another interpretation of the genericity condition is that provided $(\varepsilon, \lambda, b)$ are drawn randomly from a smooth distribution over some subset of $\mathbb{R}_+^2 \times (-1, 1)$, then with probability 1 the principal's optimal strategies will be manipulation-invariant.

[7] Note that by "directed path" we allow for the possibility that $v_i = v_j$ for $i \neq j$ along the walk (i.e., repeated vertices).

[8] By convention, any walk containing a stubborn agent will necessarily have weight zero, so taking $\mathcal{W}_{ij}$ to be the set of all walks gives identical results, but makes it easier to misapply the result (by including walks that pass through stubborn agents, and failing to zero the weight of the entire walk).

for the principal, which directly depends on the choice of interference x^*. We now define our key centrality measure:

Definition 3.1.2. The *DeGroot centrality* of agent i is equal to:

$$\mathcal{D}_i(\boldsymbol{\gamma}) = \sum_{j=1}^{n} \left(\theta_j \gamma_j \sum_{W \in \mathcal{W}_{ij}} w_W \right)$$

Our centrality measures captures the *level of influence* that other DeGroot agents have on agent i's own belief. The next proposition shows that this centrality measure, applied to any agent i is exactly equal to that agent's belief of the incorrect state:

Proposition 3.1.2. *The DeGroot centrality of agent i is equal to her limiting belief $\pi_{i,\infty}(R)$ of the incorrect state R when the principal exerts influence $\boldsymbol{\gamma}$, i.e., $\mathcal{D}(\boldsymbol{\gamma}) = (\mathbf{I} - \mathbf{W})^{-1}(\boldsymbol{\gamma} \odot \boldsymbol{\theta})$.*

Proposition 3.1.2 therefore establishes that the DeGroot centrality of agent i, defined as the weighted-sum of all stubborn-avoiding walks to other DeGroot agents, corresponds precisely to her belief in the incorrect state R, as given by Theorem 3.1.1. Moreover, the DeGroot centrality of an agent i can also be related to the centralities of her neighbors via the following recursive relationship:

$$\mathcal{D}_i(\boldsymbol{\gamma}) = \theta_i \gamma_i + \sum_{j=1}^{n} w_{ij} \mathcal{D}_j(\boldsymbol{\gamma})$$

where by definition the DeGroot centrality of a stubborn agent is 0. We provide an example of how to apply both the weighted-walk and recursive definitions of DeGroot centrality, and their equivalence to beliefs of the incorrect state in appendix A.

Discussion. As we mentioned before, our definition of centrality shares some similarities with other measures in the literature. There are three key parts of the DeGroot centrality definition: (i) longer walks are discounted more than shorter walks (i.e., closer friends have more impact than those further away), (ii) there is differentiation between agents who are influenced by the principal (targeted DeGroots) and those who are not (stubborn agents and non-targeted DeGroots), and (iii) there is a normalization of the weights so that more neighbors means less influence per neighbor (i.e., the sum of influence weights is always 1). Table 3.2 illustrate which of these properties are shared in eigenvector, Katz-Bonacich, and PageRank centrality.

Centrality	Discounted Walks	Asymmetric Influence	Normalized
Eigenvector	✗	✗	✗
Katz-Bonacich	✓	✗	✗
PageRank	✓	✗	✓
DeGroot	✓	✓	✓

Table 3.2. Comparison of Centrality Measures.

In particular, none of the centrality measures capture property (ii). This property highlights the fact that the other measures solely describe graph or network properties, whereas DeGroot centrality captures both network properties *and* the principal's strategy: DeGroot agents who are targeted by the principal contribute towards the centrality of an agent, but those who are not targeted do not. In that sense, different nodes in the network exert different types of influence on the centrality measure of other nodes, and that type is in turn dependent on the targeting strategy. For instance, PageRank centralities do not change if the network remains the same; on the other hand, if the principal plays some network strategy x_1 instead of x_2, the DeGroot centralities *will* change, even if the underlying network itself does not. Thus, the defining feature of DeGroot centrality is its ability to not only capture network structure, but also capture how the strategic provision of information shapes centrality.

Finally, we note that property (i) is also more general in DeGroot centrality than in the typical Katz-Bonacich or PageRank centrality sense: Because every node represents an agent with a heterogeneous θ_i, the discount (or dampening factor) applied in each step of the walk is different from one node to the next, and thus provides some additional subtlety which is explored in Section 3.1.4.

3.1.4 Principal's Optimal Influence

Throughout Section 3.1.3, we have characterized the learning dynamics of the population, holding fixed the influence of a possibly strategic principal. A key determinant of manipulation, however, is *how* a strategic principal chooses his influence to maximize his own payoff. Holding the network topology fixed, we consider how changes in the environment affect the principal's strategy, agents' beliefs, and the persistence of manipulation. In the next section (Section 3.1.5), we study how these elements are affected by the network topology.

We will say that society is *impervious* to manipulation if no agents are manipulated in the principal's optimal strategy; otherwise it is *susceptible*.[9] Our first comparative static analyzes the effects from changing cultural norms relating to information assimilation and social learning. The second demonstrates how influence costs and the relative payoffs from taking the incorrect vs. correct action shape the principal's intervention, sometimes in counterintuitive ways. Underlying both of these is the tension between a principal trying to spread misinformation, and stubborn agents spreading knowledge of the true state.

Comparative Statics on Personal Experience: Cultural Norms

This section examines the effect that the personal experience term θ has on manipulation. The way agents take into account their own experience relative to the opinions of others can vary substantially. An agent might put little weight on her own experience relative to what she hears from her friends (because, for example, she believes she is not well-informed about the topic at hand). Conversely, an agent might weigh her own experience much higher compared to the information she obtains from her friends, or she can simply weigh her experience similarly to her friends' beliefs. As we show, these variations lead to substantial differences when it comes to manipulation. In what follows, we study what happens for a fixed network structure as the vector of experience weights $\boldsymbol{\theta}$ changes.

Definition 3.1.3 (Network Preservation)**.** We say $(\mathbf{G}', \boldsymbol{\theta}')$ is a *network preservation* of $(\mathbf{G}, \boldsymbol{\theta})$ if $w'_{ij} = w_{ij}(1-\theta'_i)/(1-\theta_i)$ for all DeGroot agents i.

A network preservation corresponds to a shifting of weights between an agent's own experience and that of her neighbors' opinions, while preserving the relative proportions of the network weights. We call this network preservation *homogeneous* if it is a network preservation with $\boldsymbol{\theta} = \theta\mathbf{1}$ and $\boldsymbol{\theta}' = \theta'\mathbf{1}$ (i.e., all agents have the same experience weights before and after). The homogeneous network preservation corresponds to a unilateral shift in attitudes about the importance of one's own perceptions. Most naturally, in a homogeneous network, θ can be thought of an attitude parameter tuned to the cultural norms of the population.

For the following result, we fix the homogeneous network $\mathbf{G}_\theta$ with an arbitrary self-experience weight $\boldsymbol{\theta} = \theta\mathbf{1}$. For simplicity, we make the additional assumptions that in $\mathbf{G}_\theta$: (i) there exists at

[9] This implies that manipulation is a binary property of the network: it either exists or not. Section 3.1.6 extends this definition to consider *how many* agents are manipulated.

least one stubborn agent in the population, and (ii) there is at least one DeGroot not adjacent to a stubborn agent:

Theorem 3.1.3. *Let $\mathbf{G}$ be an arbitrary network with homogenous θ, and let $\mathbf{G}_{\theta'}$ denote the network preservation of $\mathbf{G}$ where all agents have θ'. There exist cutoffs $0 < \underline{\theta} < \theta^* < \overline{\theta} < 1$ such that:*

(a) If $\theta' \in (0, \underline{\theta})$, the network $\mathbf{G}_{\theta'}$ is impervious for any $\varepsilon > 0$.

(b) The network $\mathbf{G}_{\theta'}$ is impervious for $\theta' \in (\theta^*, \overline{\theta})$ only if it is impervious for $\theta' \in (\theta^*, 1)$ for any $\varepsilon > 0$.

(c) If $b > 1/2$,[10] there exists $\varepsilon > 0$ such that when $\theta' \in (\theta^*, \overline{\theta})$ the network $\mathbf{G}_{\theta'}$ is susceptible, but when $\theta \in (\overline{\theta}, 1)$ the network $\mathbf{G}_{\theta'}$ is impervious.

Theorem 3.1.3 shows that the comparative statics on manipulation are *non-monotone* in θ. A society that supports an intermediate amount of weight on each agent's own experience is the society that is most susceptible to manipulation. While social learning can be both helpful and detrimental to uncovering truth, it is most harmful (in the presence of strategic interventions) when used in moderation.

Manipulation becomes impossible when a society is more inclined towards herding (i.e., very small θ), as it relies entirely on social learning. If the community has at least one stubborn agent, then the beliefs of that agent spread throughout the network. This may come at the cost of agents dismissing accurate information from organic news sources and thus learning more slowly, but guarantees agents will eventually find the truth. Conversely, social learning plays little role in a culture that supports strong individuality and narcissism (i.e., very large θ). Thus, the principal cannot exploit social network effects to propagate his message, i.e., the principal is no longer able to reach a large population by only targeting a small subset of agents, and instead has to reach all agents directly (e.g., door-to-door campaigning), which is costly. With intermediate θ, however, agents both incorporate their own experience *and* employ social learning, allowing the principal to leverage social forces to spread his message without getting completely drowned out by the stubborn agents.

[10] When b is small, the network can exhibit no manipulation for any θ' or a "phase transition" instead: there exists θ^{**} such that $\theta' < \theta^{**}$ is impervious but $\theta' > \theta^{**}$ is susceptible.

The next result considers heterogeneous settings and stands in contrast to Theorem 3.1.3. Let us consider a set D_1 of DeGroot agents with θ_1 and a set D_2 of DeGroot agents with θ_2.

Proposition 3.1.3. *Suppose agents in D_1 are strongly connected and there exists at least one link from D_2 to D_1. For fixed θ_2, there exists $\bar{b}$ such that for all $b > \bar{b}$, even as $\theta_1 \to 0$, all DeGroot agents (including those in D_1) are manipulated for sufficiently small ε. On the other hand, if $\theta_1 = \theta_2 = \theta$, for every $b < 1$ there exists $\bar{\theta}$ such that for all $\theta < \bar{\theta}$, the network is impervious if there is at least one stubborn agent in the network, regardless of ε.*

In heterogeneous settings (where there might be more than one value of θ in society), even if some agents have a small θ and discount all news from the principal, they can still mislearn the state if other agents hold high θ. Those agents discounting their own experiences are now manipulated because they listen mostly to the experiences of others, who may be voicing the beliefs of stubborn agents, but may also be voicing the misinformation they receive from the principal.

Network Formation Considerations. The above observations highlight a novel channel for the formation of social networks as a means for avoiding misinformation. We briefly detour to consider how agents in a society might choose the weights to assign to their personal experiences so as to maximize their chances of learning the correct state. Formally, if an agent learning through DeGroot-style heuristics could choose her θ_i, how should she do so?

Note that if other agents are relying strongly on social learning over personal experience (i.e., low θ_j), then agent i can benefit by setting $\theta_i \approx 0$ as well to receive influence from only those agents who know the truth with certainty. However, as other agents increase their θ_j, agent i will conform to the beliefs of her immediate peers who, as observed in Proposition 3.1.3, may or may not be more amenable to misinformation. In a more individualistic culture, if agent i believes she is more discerning of the news relative to her peers, she is better off picking a larger θ_i herself when the values of θ_j are large.

This defines a coordination game where agents try to arrive at a cultural standard for θ by matching others' choices. When agent i does not match this cultural norm, she risks making a naive decision while ignoring (smart) stubborn agents in the population (picking θ_i high when others pick low) or risks listening to bad advice when knowing better herself (picking θ_i low when others pick high). Loosely, the equilibria of this game correspond to the entire spectrum of homogenous θ. But as we saw in Theorem 3.1.3, some equilibria can be more socially inefficient than others. For instance, when agents choose an intermediate θ that splits learning between personal experience and social forces, manipulation is generally worst for society.

Influence Costs and Payoff Asymmetry

We conclude this section by considering how manipulation is affected by the cost of sending misinformation and the payoff asymmetry under the two different actions S and R. Recall that ε captures the per-agent cost of sending misinformation,[11] whereas b parametrizes the agent's natural affinity toward one action or the other.[12] We obtain the following comparative static:

Proposition 3.1.4. *When ε increases, the number of manipulated agents never increases (but may decrease). Similarly, if the network is susceptible with payoff asymmetry b, it is still susceptible when increasing b.*

Perhaps counterintuitively, however, increases in b can create incentives for the principal to target fewer agents and decrease manipulation as a whole, as seen in the following example:

Example 3.1.1 (Targeting "Low Hanging" Fruit)**.** Consider the social network consisting of three DeGroots and one stubborn agent arranged along a bidirectional line as in Figure 3-1. Let $\varepsilon \in (1, 3/2)$ throughout. All DeGroot agents weigh their neighbors and themselves according to $\theta_i = \alpha_{ij} = 1/(1 + |N(i)|)$. First, suppose that $b = 0.3$, so the belief cutoff to take action R is given by $\pi_{\text{cutoff}}(R) = 0.35$. Then it can be shown that the principal manipulates all three agents by sending misinformation to agents 1 and 3, which yields payoff $3 - 2\varepsilon > 0$.[13] Now suppose b increases to $b = 0.6$, so the belief cutoff to take action R is given by only $\pi_{\text{cutoff}}(R) = 0.2$. Then, one can show the principal manipulates only two of the three agents by sending misinformation to only one agent (for instance, by manipulating agents 1 and 2 through sending signals to agent 1), which yields a payoff of $2 - \varepsilon > 3 - 2\varepsilon$. Thus, while manipulation became "easier" because the cutoff required to take the wrong action had decreased, the number of manipulated agents also decreases from 3 to 2. ∎

[11] More general cost structures are considered in Section 3.1.6.

[12] For instance, if consuming a risky product provides more disutility than a safe product provides utility, then $b > 0$, indicating that the agent must be (much) more confident in the product's safety to choose action S.

[13] To see this, observe that sending more signals cannot improve the principal's payoff, and targeting only a single agent leads to only that agent being manipulated when $b = 0.3$, which is worse than targeting no one since $\varepsilon > 1$. A more formal proof is given in Appendix B.1.2.

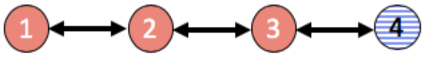


Figure 3-1. Network for Example 3.1.1 (Solid = DeGroot, Shaded = Stubborn).

3.1.5 Network Topology

We now turn our attention to examine how the topology of the network and the placement of stubborn agents affect manipulation. We provide a classification of networks into dense and sparse topologies, and compare these structures in terms of what is required to make them impervious. The upshot is that a *constant* number of these stubborn agents located *anywhere* in a dense network makes it impervious. Sparse networks on the other hand require more resources (number of stubborn agents needed) and planning (where to place these agents). Even when the number of stubborn agents is enough to make the network impervious, the location of these agents have to be carefully chosen. Further, it is possible that there are scenarios where the number of stubborn agents required grows with the size of the network, making imperviousness more difficult to achieve.

Dense Networks

We start by defining what it means for a network to be dense. Recall that $\mathcal{W}_{ij}$ is the set of all stubborn-avoiding walks between i and j. To this end, we define the *log-diameter* of the network $\mathbf{G}$ as:

$$d_{\mathbf{G}} \equiv \max_{i,j} \min_{W \in \mathcal{W}_{ij}} \sum_{(k \to \ell) \in W} -\log(w_{k\ell})$$

where the weights w_{ij} are from the matrix-representation $\mathbf{W}$ of $\mathbf{G}$ (see Section 3.1.2). Using this, we can define the *density* of a network as follows:

Definition 3.1.4 (Dense Networks)**.** We say that network $\mathbf{G}$ is δ-*dense* if it has a log-diameter of at most $\log(n + \delta)$.

We can then utilize this definition to give the following result:

Theorem 3.1.4 (Constant Number of Stubborn Agents)**.** *For every δ, there exists a universal constant $m^*(\delta)$ such that every network $\mathbf{G}$ which is δ-dense and contains at least $m^*(\delta)$ stubborn* *agents is impervious to manipulation.*[14]

Theorem 3.1.4 implies that a vanishingly small fraction of stubborn agents in the population is all that is required to make the principal unable to manipulate beliefs. Moreover, if the positions of those stubborn agents were to be chosen by an adversary, manipulation will still not be possible as long as the network G satisfies the log-diameter condition for every placement of $m^*(\delta)$ stubborn agents. Finally, note that the shortest path between agent i and every stubborn agent being less than $\log(n+\delta)$ does not automatically imply that agent i will not be manipulated. This needs to hold *uniformly* across all DeGroot agents, otherwise the log-diameter condition is not satisfied. Example 4 in Appendix B.3.2 demonstrates how a DeGroot agent that is directly connected to a stubborn agent can still believe the incorrect state as a result of her living in a DeGroot bubble where echo chamber effects are rampant.

Theorem 3.1.4 guarantees that if the number of stubborn agents meets the threshold $m^*(\delta)$ in a δ-dense network then there will never be manipulation, but this bound may not be tight. In Appendix B.1.4, we perform a numerical study of how the number of stubborn agents and their placements might affect manipulation in practice, as compared to the log-diameter bound provided in Theorem 3.1.4.

We conclude this section by mentioning a few examples of interest where one can easily apply the result of Theorem 3.1.4, along with one cautionary example where the result cannot be utilized. These examples are worked out in detail in Appendix B.1.3.

(i) *The complete network*: The complete network is the most dense network, and is impervious provided the number of stubborn agents satisfies $m \geq (1+b)/(1-b)$.

(ii) *Influential star network*: In the influential star network, most agents listen to a single (central) agent. This network can be impervious even if the central agent herself is DeGroot. This occurs because the network is sufficiently dense, as it is possible to get from one agent to another by passing through that central agent. We can then apply Theorem 3.1.4 and show that $m \geq 2(1+b)/(1-b)$ stubborn agents are sufficient for imperviousness, irrespective of their location.

(iii) *Echo chamber network*: An echo-chamber network is a network where DeGroot agents communicate almost-exclusively with other DeGroot agents. For instance, consider two

[14] For the interested reader, Appendix B.1.1 offers a stronger version of Theorem 3.1.4 that requires satisfying a weaker notion of local density everywhere in the network.

cliques of size $n/2$, one of all DeGroots and one of all stubborn agents, with a single connection between them. The unweighted network G^* has diameter 3 for all n, but admits the same manipulation as in Example B.1.4, despite a *linear* number of stubborn agents. It is easy to check the shortest path between most DeGroots and a stubborn agent is roughly $\log(n^2/2)$, so does not satisfy the log-diameter condition of Theorem 3.1.4 for any δ.

Susceptible Networks: An Example

As a prelude to our discussion of sparse networks, we demonstrate the traits that make such networks susceptible to manipulation by considering the directed ring network as a prototypical example. We follow this up with a more general characterization in Section 3.1.5.

In an episode of the show Planet Money on NPR, the political consultant David Goldstein discusses how firms like Cambridge Analytica interfere in elections by targeting agents with messages in order to push them towards specific actions, and how this strategy can be profitable even if it fails to sway most agents who receive such messages:[15]

> "You might be immune and the guy next to you might be immune, and the guy next to that person might be immune, but if I only need to change [the minds of] 3% of people in order to affect a given result, then I can go 97 people down and not have an effect but as long as I have an effect on 1, 2, and 3, then I can literally change the world."

Goldstein's quote was made in a different context from the one we consider, but it perfectly encapsulates the example of the ring network in Figure 3-2. In this example, the principal targets several agents with misinformation despite knowing that some of these agents will not be directly affected and will still figure out the correct state of the world. This targeting however reverberates through the network, and allows the principal to manipulate agents at the end of the ring without sending them any messages, leading to an overall lower cost of manipulation. We now discuss this example in detail.

Consider a ring network with homogenous $\theta_i = \theta$. Network weights are given by $\alpha_{ij} = 1 - \theta$ for $j = i - 1$, and $\alpha_{ij} = 0$ for all other j. Under this assumption, each DeGroot listens to her

[15] This quote comes at the 16:30 minute mark at https://www.npr.org/2019/05/24/726536757/episode-915-how-to-meddle-in-an-election

own news and the opinion of her immediate neighbor, who in turn listens to her immediate neighbor, etc. Consider the following stubborn agent placements:

1. *Continuous chain*: Assume the first m agents on the ring network are all stubborn (i.e., the stubborn agents talk *mostly* with other stubborn agents), and the remaining agents are DeGroots.

2. *Sprinkled*: The stubborn agents are "sprinkled" throughout the network so that the distance of any DeGroot agent i to the nearest stubborn agent is minimized.

Continuous Stubborn Agent Chain. For illustration, consider the case of a continuous chain of stubborn agents with $\theta = \frac{1}{n+1}$ and $m \ll n$, so that there are fewer stubborn agents than DeGroots. Suppose the principal solves an easier influence problem along only a single dimension: (i) he exerts influence along a continuous arc in the ring, and then does not exert influence for the remaining agents, and (ii) he wants to induce the maximal number of manipulated agents. This is not necessarily the principal's optimal network strategy, but we use this to show that there is *some* strategy that beats $\mathbf{x} = \mathbf{0}$, and therefore there must be some manipulation by Corollary B.1.1. An illustration of this strategy is given in Figure 3-2.

Consider DeGroot agent i at location τ away from the last stubborn agent. We can write her belief in terms of her DeGroot centrality $\mathcal{D}_i(\boldsymbol{\gamma})$, a function of $\boldsymbol{\gamma}$:

$$\mathcal{D}_i(\boldsymbol{\gamma}) \sim \sum_{j=0}^{\tau-1} \frac{n^j}{(n+1)^{j+1}} \gamma_{\tau-j}$$

If the principal has chosen $\gamma_i = 1$ for all previous agents, then the above reduces to:

$$\mathcal{D}_i(\boldsymbol{\gamma}) \sim 1 - \left(\frac{n}{n+1}\right)^{\tau}$$

when τ is sublinear, $\mathcal{D}_i(\boldsymbol{\gamma}) \to 0$, whereas when $\tau = \alpha n$, we get $\mathcal{D}_i(\boldsymbol{\gamma}) \to 1 - e^{-\alpha}$. Recalling that agents with $\mathcal{D}_i(\boldsymbol{\gamma}) > (1-b)/2$ will choose the incorrect action, we find that all but $\log\left(\frac{2}{1+b}\right)$ proportion of the DeGroot agents are manipulated when $b \geq (2-e)/e$. Therefore, *even with a growing population of stubborn agents, a linear number of DeGroot agents are manipulated.*

When stubborn agents form a continuous chain, the network is fundamentally equivalent to one where the chain is replaced by a single stubborn agent who knows the truth at the end

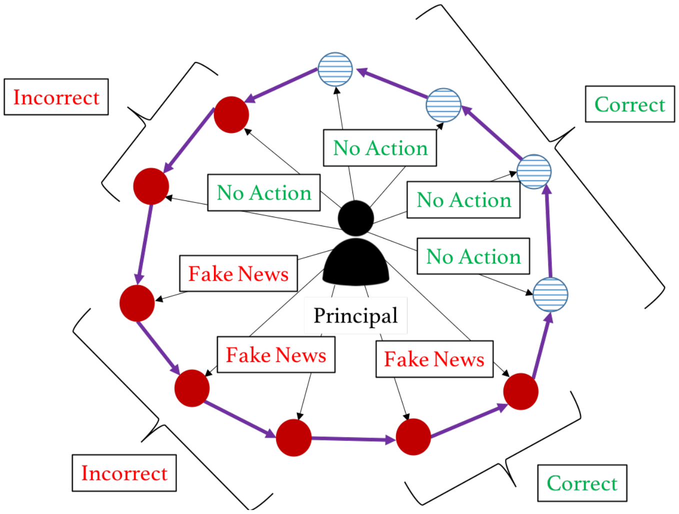


Figure 3-2. Beliefs in the Ring Network. A directed arrow from node i to node j indicates that i listens to j. Shaded nodes represent stubborn agents.

of this chain. The long chain of DeGroot agents who receive misinformation drowns out the beliefs of the DeGroots at the beginning of the ring.

Sprinkled Stubborn Agents. We now consider the effects of stubborn agent placement by "sprinkling" them throughout the ring. Unlike with the continuous chain, in this case, we obtain the following result:

Proposition 3.1.5. *There exists a constant m^* such that if there are $m > m^*$* sprinkled *stubborn agents in the ring network, it is impervious to manipulation for any n.*

In contrast to Theorem 3.1.4, Proposition 3.1.5 imposes firm restrictions on the placement of stubborn agents, but similar to that theorem, it shows that only a constant number are needed. Recall in dense networks, neither the *placement* nor *the number* of stubborn agents have to meet particularly demanding conditions to guarantee imperviousness. However, in the ring network, placement becomes crucial, despite still only requiring a small fraction of stubborn agents to avoid manipulation. Although the ring network is sparse, because agents largely discount their own experiences ($\theta \sim 1/n$) and echo chambers are limited,[16] a few optimally-placed stubborn agents limit the spread of misinformation.

Recall that because $\theta = 1/(n+1)$, as the network gets large, agents mostly dismiss their

[16] Because opinions only flow in one direction, echo chambers are not too strong. Here, sparsity is the main driver of manipulation, and thus requires special placement to avoid it. In the case of other sparse networks, such as the bidirectional ring or cliques of all DeGroots (e.g., Example B.1.4), echo chambers can be much worse and drive beliefs even farther away from truth.

own experiences. A more natural assumption is to suppose agents weigh all social influences equally with their own experience. Formally, we consider the *equal-influence* weighting given by:

$$\theta_i = \alpha_{ij} = \frac{1}{1 + |N(i)|} \tag{3.3}$$

for DeGroots i and all $j \in N(i)$. While in the complete network this corresponds to setting $\theta_i = 1/(n+1)$ as before, in the ring this instead admits weighting $\theta_i = \theta = \alpha_{i(i-1)} = 1/2$. In contrast to Proposition 3.1.5, when DeGroots listen more to their own news, we obtain a result in stark contrast to the case of dense networks:

Proposition 3.1.6. *Consider the ring network with equal-influence weighting and $\varepsilon < 1$. Then there exists a constant $c > 0$,*[17] *such that the network is impervious with $c \cdot n$ sprinkled stubborn agents agents, but is susceptible if there are fewer than $c \cdot n$ stubborn agents* or *their configuration is not sprinkled.*

In the equal-influence ring network, imperviousness comes with stringent requirements on *both* resources (number of stubborn agents) and planning (their placement). First, as the network grows in size, the number of stubborn agents must also grow in proportion, so that a constant fraction of the population is still stubborn. Second, the location of the stubborn agents is paramount to preventing manipulation. We summarize these findings in Table 3.3.

Network	Resources	Planning	# Manipulated (when possible)
Dense Network	$\Theta(1)$	Anywhere	$\Omega(1)$
Ring Network	$\Theta(1)$	Sprinkled	$\Omega(n)$
Equal-Influence Ring	$\Theta(n)$	Sprinkled	$\Omega(n)$

Table 3.3. Properties based on Network Density.

The principal's ability to manipulate in networks that do not satisfy the log-diameter condition of Theorem 3.1.4 is not unique to the ring network. In addition to the more general characterization of sparse networks in the next section, Example B.1.6 in Appendix B.1.3 provides another demonstration of susceptibility on the star network with equal-influence weighting.

[17] Here, and throughout the entire paper, by *constant* we mean $c \in \Theta(1)$, so there exist $\underline{\beta}, \overline{\beta}$ independent of n such that $\underline{\beta} \leq c \leq \overline{\beta}$.

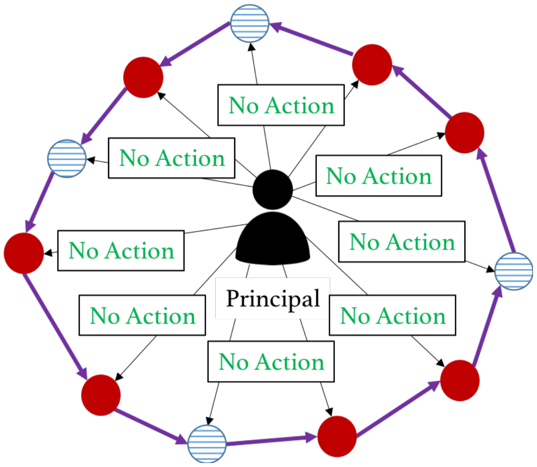


Figure 3-3. An illustration of Proposition 3.1.5 with "sprinkled" stubborn agents.

Sparse Networks

We now generalize the insights of Section 3.1.5 to a wide array of sparse networks. First, we consider a continuum of networks that are parametrized by a sparsity parameter η, and proceed to give a characterization of manipulation for all η. Second, we provide a sufficient condition for imperviousness in symmetric networks. Because symmetric networks may either be dense (e.g., complete) or sparse (e.g., ring), this result provides a more complete understanding of imperviousness across different levels of sparsity. Our main findings highlight how the requirements on resources and planning become more demanding as the network becomes more sparse, corroborating the results from Section 3.1.5.

Convex Combination of Ring and Complete Networks Let us fix the stubborn and DeGroot agents in the population and consider two different network structures $\mathbf{G}_c = (\boldsymbol{\theta}^c, \mathbf{W}^c)$ and $\mathbf{G}_r = (\boldsymbol{\theta}^r, \mathbf{W}^r)$, corresponding to the complete network and the ring network with equal-influence weighting. By Theorem 3.1.4 and Example B.1.2, we know $\mathbf{G}_c$ is impervious to manipulation with a constant number of stubborn agents (located anywhere). On the other hand, in $\mathbf{G}_r$, an unbounded (in n) number of agents are manipulated whenever the number of stubborn agents is sublinear *or* these agents are not in specific network positions.

Now consider parameter $\eta \in [0,1]$ and define the network $\mathbf{G}_\eta$ as $\theta_i = \eta \cdot \theta_i^c + (1-\eta) \cdot \theta_i^r$, and $\alpha_{ij} = \eta \cdot \alpha_{ij}^d + (1-\eta) \cdot \alpha_{ij}^s$ for all i, j. Note that as η varies from 0 to 1, the network becomes more dense and, by construction, the network is susceptible at $\eta = 0$ but impervious at $\eta = 1$. The following result provides the full characterization for intermediate η:

Theorem 3.1.5. *Suppose there are either $o(n)$ stubborn agents or that these agents form a continuous chain in the ring. There exists η^* such that: (i) if $\eta < \eta^*$, $\mathbf{G}_\eta$ the number of manipulated agents grows unboundedly in the size of the network n, whereas (ii) if $\eta > \eta^*$, $\mathbf{G}_\eta$ is impervious to manipulation.*

Theorem 3.1.5 shows that a *phase transition* exists between dense and sparse networks: when the network gets sufficiently dense, all opportunities for manipulation suddenly vanish. Similarly, it shows the results of Section 3.1.5 are fairly robust: qualitative conclusions for the ring network generalize to sparse networks even as they become slightly more dense.

Symmetric Networks of Degree k We now consider *symmetric networks*, where any two agents have identical network positions. In particular, for any directed unweighted network $\mathbf{G}^*$, we say $\mathbf{G}^*$ is symmetric if and only if for every pair of vertices i, j, there exists a function $f : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $f(i) = j$ and $k \to \ell$ exists in $\mathbf{G}^*$ if and only if $f(k) \to f(\ell)$ exists in $\mathbf{G}^*$.[18] We say a network $\mathbf{G}$ is *symmetric* if the unweighted analog, $\mathbf{G}^*$, is symmetric and $\{\alpha_{ij}, \theta_i\}$ satisfy the equal-influence weighting (Equation 3.3) for all links $i \to j$ that exist in $\mathbf{G}^*$ (i.e., $j \in N(i)$ whenever $i \to j$ exists in $\mathbf{G}^*$, and $\alpha_{ij} = 0$ if $j \notin N(i)$).

In other words, a symmetric network $\mathbf{G}$ is one where all agents are symmetric in the unweighted sense, and employ equal-influence weighting. When the network is strongly connected, k-regularity (i.e., each agent has k neighbors) is a necessary (but not sufficient) condition for symmetry, so in particular we have $\theta_i = \alpha_{ij} = 1/(1+k)$ for all agents i whenever there exists a link $i \to j$. Therefore, symmetric networks can be partitioned into classes of "degree-k" symmetric networks. We will also say K is a *symmetric placement* of stubborn agents if the induced subgraph $\mathbf{G}^* \backslash K$ is symmetric.

Suppose that a fraction ϕ of all the links going into stubborn agents are links between stubborn and DeGroot agents. Then, within the class of symmetric networks, we obtain the following characterization:

Theorem 3.1.6. *Suppose $\mathbf{G}$ is a degree-k symmetric network with a symmetric placement of $m = |K|$ stubborn agents. Then the network is impervious to manipulation if $\phi k m/(n-m) = \phi k |K|/|D| \geq (1+b)/(1-b)$.*

Theorem 3.1.6 further demonstrates how sparsity tends to make manipulation easier.

[18] This is simply the definition of a graph automorphism.

Here the degree of the agents, k, functions as a measure of sparsity, and along with the ratio of stubborn agents to DeGroots in the population, $m/(n-m) = |K|/|D|$, determines a sufficiency condition for imperviousness. Moreover, Theorem 3.1.6 highlights how stubborn agent placement becomes more demanding with sparsity, on two fronts: first, the placement must be *symmetric*, and cannot be arbitrarily chosen, and second, the stubborn agents ought to be placed in such a way that the links going from these agents to DeGroots (and vice-versa) are maximized. Both of these requirements become more difficult as the network becomes sparser.

The bound in Theorem 3.1.6 is in fact tight in many common network topologies. In the case of the complete network, we have $\phi k = |D|$, so $m \geq (1+b)/(1-b)$, which is the exact bound we saw in Example B.1.2. In this case, any placement of stubborn agents is symmetric and has the same ϕ, so the restriction in Theorem 3.1.6 is immaterial. The result is also tight in the ring network, where $k = 1$, so m needs to be linear in n to avoid manipulation (as in Proposition 3.1.6). Here, the symmetric placement is more challenging and requires careful planning; unsurprisingly, the symmetric placement corresponds precisely to the "sprinkling" arrangement of Section 3.1.5.

3.1.6 Extensions

In an effort to illustrate how social learning changes in the presence of strategic interventions, we have presented a parsimonious framework with a number of simplifying assumptions. In this section, we consider how the results and conclusions change in the face of additional complications. While these extensions offer further areas of exploration and more detailed analyses, they also demonstrate how our simplified framework can be applied without much loss of generality.

Learning the Principal's Type

In Section 3.1.2, we have assumed that agents share their beliefs about the state with their neighbors, but not about the type of the principal. We now endow the DeGroot agents with some degree of skepticism. Agents are aware of the possibility of a strategic principal, and in addition to learning about the state, they also update their beliefs on whether the news they receive is organic or strategic. Does this skepticism always decrease manipulation?

To provide an answer to this, we introduce a coupled belief dynamics process for DeGroot agent who may be aware of possible misinformation. In addition to sharing beliefs $\pi_{i,t}$ about the state, we assume agents also share $\mu_{i,t}$, their belief that the principal is truthful instead of strategic. Every DeGroot agent has prior $\mu_{i,0}$ about whether their news source is entirely organic, and personal-experience weight $\tilde{\theta}_i$ about this prior. Moreover, we assume that DeGroot agents exchange beliefs about the principal's type according to the influence matrix $\tilde{\mathbf{W}}$ (where $\tilde{\mathbf{W}}$ is not restricted to be equal to $\mathbf{W}$).

The coupled dynamics process occurs as follows. Agents endogenously choose how much weight to put on the belief they form from reading the news. This weight is directly proportional to how trustworthy they believe the news source is. If an agent believes much of the news they receive is misinformation sent by the principal, then the agent puts much more weight on social learning and largely dismisses the news she observes. Thus, instead of putting a constant weight θ_i on their own news, DeGroot agents put $\mu_{i,t}\theta_i$ weight on their personal news update. Formally, the belief update process obeys the following law of motion:

$$\mu_{i,t+1} = \tilde{\theta}_i \mu_{i,0} + \sum_{j=1}^{n} \tilde{\alpha}_{ij} \mu_{j,t}$$

$$\pi_{i,t+1} = \mu_{i,t}\theta_i \cdot \mathrm{BU}(h_{i,t}) + \frac{1 - \theta_i \mu_{i,t}}{1 - \theta_i} \sum_{j=1}^{n} \alpha_{ij} \pi_{j,t}$$

Note that as $\mu_{i,t} \to 1$, we recover the baseline model from Section 3.1.2, whereas when $\mu_{i,t} \to 0$, agents dismiss their personal experience entirely. With this formulation, the next result shows how we can reduce this belief process to the baseline model:

Proposition 3.1.7. *The coupled-belief dynamics process is equivalent to the baseline model where agents use personal-experience weights $\boldsymbol{\theta}'$ given by:*

$$\boldsymbol{\theta}' = \boldsymbol{\theta} \odot (\mathbf{I} - \tilde{\mathbf{W}})^{-1} (\boldsymbol{\mu}_0 \odot \tilde{\boldsymbol{\theta}})$$

and the corresponding network preservation on $\mathbf{W}$.

Proposition 3.1.7 shows how the personal weights of the belief update process can arise endogenously when agents engage in a coupled belief update that considers the trustworthiness of their own news. One can apply similar comparative statics as in Section 3.1.4 to understand how DeGroot skepticism affects limit beliefs about the state:

(a) Uniformly more skepticism about the veracity of information (i.e., lower $\boldsymbol{\mu}_0$) does not necessarily lead to better outcomes (i.e., less manipulation): $\boldsymbol{\theta}'$ is increasing in $\tilde{\boldsymbol{\theta}}$, and by Theorem 3.1.3(c), it is possible for manipulation to *increase* when $\boldsymbol{\theta}$ decreases. That being said, sufficient skepticism across all agents leads to imperviousness by Theorem 3.1.3(a), when there is at least one stubborn agent in the population.

(b) For the same reason as (a), the baseline model may protect more agents from manipulation than this revised model where agents take into account the possibility of misinformation.

(c) Extreme skepticism (as opposed to just *more* skepticism) about the veracity of information does not necessarily protect a given agent. By Proposition 3.1.3, if other agents are less skeptical, then this agent can still be manipulated by absorbing misinformation acquired from social learning.

(d) However, when $\varepsilon \approx 0$, additional skepticism about the accuracy of news always improves the beliefs of DeGroot agents. These can be seen directly through the DeGroot centrality expression, $(\mathbf{I} - \mathbf{W})^{-1}(\mathbf{1}_D \odot \boldsymbol{\theta})$, and given that $(\mathbf{I} - \mathbf{W})^{-1}$ contains all non-negative entries, it is monotone in $\boldsymbol{\theta}$.

Alternative Cost Functions

We have assumed throughout that the principal's cost function follows the form $c(\mathbf{x}) = \sum_{i=1}^{n} \varepsilon x_i$. In particular, we have assumed costs are *linear* and *homogenous* across agents. We consider two variants of this:

1. *Non-linear specification*: Suppose that $c(\mathbf{x}) = C\left(\sum_{i=1}^{n} x_i\right)$, but that c may not scale directly with $X \equiv \sum_{i=1}^{n} x_i$. In particular, there may be *concave costs* with the intervention: we assume $C' > 0$ but $C'' < 0$, with $C'(0) > \varepsilon$ and $\lim_{X\to\infty} C'(X) = 0$. Similarly, there may be *convex costs* with the intervention: we assume $C' > 0$ and $C'' > 0$, with $C'(0) < \varepsilon$ and $\lim_{X\to\infty} C'(X) = \infty$.

2. *Heterogenous costs*: Certain agents may be more expensive to target than others, such as celebrities or those who do not use social media, etc. Thus, we assume there is a vector of costs $\boldsymbol{\varepsilon} = \{\varepsilon_i\}_{i=1}^{n}$ so that $c(\mathbf{x}) = \boldsymbol{\varepsilon} \bullet \mathbf{x}$.

A number of results are completely unaffected by these changes, for example, Theorem 3.1.4 for dense networks (Section 3.1.5), the characterization of manipulation in symmetric networks (Section 3.1.5), and the comparative statics on θ (Section 3.1.4). This follows from the fact these sufficiency conditions establish an upper bound on the DeGroot centralities of the agents in the population, and thus hold independently of the principal's costs for intervention.

More generally, however, changes in the cost function will change the scope of manipulation (e.g., in the ring). Let $\mathbf{x}^*$ be the optimal principal intervention with cost function $c(\mathbf{x}) = \sum_{i=1}^n \varepsilon x_i$ and $X^* = \sum_{i=1}^n x_i^*$. We provide the following comparative result relating these more general cost functions to the one provided in Section 3.1.2:

Proposition 3.1.8. *Let $\bar{X}$ denote the (unique) crossing point of $C(X)$ and εX. If there is concave cost and $X^* \geq \bar{X}$ then manipulation never decreases; if there is convex cost and $X^* \geq \bar{X}$, then manipulation never increases, whereas if $0 < X^* < \bar{X}$, the network is always susceptible.*

The intuition for the result can be seen in Figure 3-4. When $X^* \geq \bar{X}$, concave costs encourage the principal to expend more resources at lower marginal cost than in the linear case, increasing manipulation; on the other hand, convex costs entice the principal to slow her influence and save on higher marginal costs (similar to Example 3.1.1).

Lastly, we note the optimization problem admitting an exact characterization of the optimal strategy (given in Appendix B.1.1) can be modified easily to account for heterogeneous ε_i without affecting the nature of the problem. In stylized examples such as the ring, star, or complete network, the analysis can be applied as is by considering the average costs of sending signals (i.e., $\frac{1}{n-m} \sum_{i=m+1}^n \varepsilon_i$) in place of ε. As such, this generalization does not affect the qualitative findings of this paper.

Extent of Manipulation

Throughout the paper, we have focused on conditions where manipulation occurs for at least one agent (or none at all). In many contexts, a more appropriate metric is the number of manipulated agents, possibly relative to the population size. While we provide some characterization of the number of manipulated agents throughout (see Table 3.3), we present here a technical reduction that shows how imperviousness can be easily generalized to this problem.

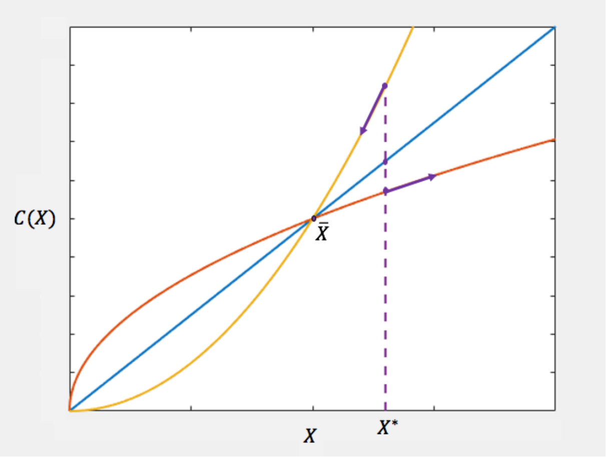


Figure 3-4. Concave vs. Convex Costs.

Definition 3.1.5. We say a network is k-*impervious* if there are k or fewer manipulated agents. Similarly, a k-*cut subnetwork* of a network G is a network obtained from coalescing a set $\mathcal{K}$ of (at most) k vertices from the network (i.e., replacing all vertices in $\mathcal{K}$ by a single vertex u) and setting $\alpha_{iu} = \sum_{j \in \mathcal{K}} \alpha_{ij}$ for all agents i, with $\theta_u = 1$.

With this transformation, we get the following reduction:

Proposition 3.1.9. *If there exists a k-cut subnetwork that is impervious to manipulation (with the exception of u) when $\varepsilon_u = 0$ (and $\varepsilon_i = \varepsilon$ for all other agents), then the original network is k-impervious.*

Therefore, having a complete understanding of k-imperviousness in networks reduces to understanding imperviousness and finding "clever" k-cuts. As an immediate corollary to Proposition 3.1.9, we get a log-diameter condition that generalizes Theorem 3.1.4 for k-imperviousness:

Corollary 3.1.1. *Consider a k-cut subnetwork with the k-cut vertex u removed.*[19] *If the log-diameter of this network does not exceed $\log(n-k+\delta)$, the network is k-impervious if it has $m > m^*(\delta)$ stubborn agents (where $m^*(\delta)$ is the same as in Theorem 3.1.4.)*

In Example B.1.7 of Appendix B.3.2, we show how in a core-periphery network, Theorem 3.1.4 cannot be applied for any value of δ. Yet, Corollary 3.1.1 establishes the network is k-impervious

[19] Note that this network is not a "valid" subnetwork in the sense that some agents have $\theta_i + \sum_j \alpha_{ij} < 1$ after the removal of u. However, DeGroot centrality (and log-diameter) are still well-defined provided that sub-stochasticity is satisfied: $\boldsymbol{\theta} + \mathbf{W}\mathbf{1} \leq \mathbf{1}$.

for a small value of k agents (agents on the periphery), so a vanishing fraction of agents are manipulated.

Dynamic Targeting Policies

In Section 3.1.2, we assume the principal chooses to send each agent with $x_i = 1$ a signal of intensity λ^*, with message $\hat{y}$, which costs him ε. Here, we relax this specification in the following ways:

(a) The principal may send different messages (call these messages $\hat{y}_i \in \{S, R\}$) and/or apply different intensities to different agents, i.e., λ_i^*.

(b) The principal may vary its message and/or intensity throughout time, i.e., $\lambda_i^*(t)$, $\hat{y}_i(t)$.

(c) The principal pays a larger cost for greater intensity messages; that is, the principal pays $\frac{1}{t}\int_0^t \tilde{\varepsilon}(\lambda_i^*(t'))\, dt'$, where $\tilde{\varepsilon}$ is an increasing, convex, and continuous function.

While this relaxation provides many more decision variables for the principal, the outcomes can be analyzed in nearly the exact same way. The following result makes that clear:

Proposition 3.1.10. *Consider the model of Section 3.1.6 with heterogenous (but linear) costs $\varepsilon_j = \tilde{\varepsilon}(\lambda(2p_j - 1))$ for each agent j. Every agent manipulated in this model is manipulated when the principal is allowed to use dynamic targeting policies, and vice-versa.*

The result of Proposition 3.1.10 should not be interpreted as dynamic targeting policies not helping the principal, but rather, that the problem can be analyzed in a static setting using an alternative cost formulation. In this setting, we see that the principal must pay higher costs to send signals to agents whose organic signals are more informative, and that those who are more skilled at interpreting organic news are more difficult to manipulate through their direct personal experience, unlike in the baseline model.

3.1.7 Conclusion

In this paper, we consider a classic social learning setup when some of the information in the network is injected by a strategic principal, and we identify conditions that allow this principal to interfere with the learning process of the agents in order to shape their beliefs. These interactions are common in marketing, public health, politics, and many other contexts,[20] and we provide a model that allows us to study them in a formal setup. We employ a diverse population that possess different degrees of knowledge about the state, which we model by using classical DeGroot agents and knowledgeable stubborn agents. We find that in this setup, the ability of a self-interested principal to manipulate a population depends on the network structure and the social norms in the network (as modeled by how much agents are willing to incorporate their friends' opinions into their own beliefs). We show that manipulation or lack thereof can be quite sensitive to these factors. In particular, we develop a centrality measure that we call DeGroot Centrality, which can be used to quickly identify which agents in the population are at risk of being manipulated. We demonstrate the use of this measure by studying manipulation in several common network topologies, and show that sparse topologies are typically more susceptible than dense ones. We demonstrate how some networks can be resilient with the presence of a small number of these stubborn agents, whereas others continue to be susceptible to manipulation unless the *number* and *location* of these agents meet certain demanding criteria.

Our work can be extended on several fronts. We have studied the dynamics of our learning model in the limit, and characterizing the strategies played by the principal in the short-term is also an important but challenging problem. Relatedly, when agents have imperfect recall (e.g. because of costly information acquisition as in Liu (2011) or recency bias, these short-term dynamics become especially relevant, even when the learning horizon is long. Finally, as discussed in Section 3.1.4, agents can use their social network as a way to protect themselves against potential misinformation. Understanding how agents form their social circles to acquire accurate information is an unexplored avenue for models of social network formation in the presence of misinformation, and provides yet another area of potential future work.

3.2 Social Inequality and Misinformation

Using the model presented in Section 3.1, we study the spread of misinformation in a social network characterized by unequal access to learning resources, based on the work of Mostagir

[20] For example, Allon and Zhang (2017) examine a model where agents learn about service quality from their experience as well as what they hear from their friends, and ask how the firm should incorporate this learning process into its decisions about which service levels to offer.

and Siderius (2022b). Agents use social learning to uncover an unknown state of the world, and a principal strategically injects misinformation into the network to distort this learning process. A subset of agents throughout the network is endowed with knowledge of the true state. This gives rise to a natural definition of inequality: *privileged* communities have unrestricted access to these agents while *marginalized* communities do not. We show that the role that this inequality plays in the spread of misinformation is highly complex. For instance, communities who hoard resources and deny them to the larger population can end up exposing themselves to *more* misinformation. On the other hand, while more inequality generally leads to worse outcomes, the prevalence of misinformation in society is non-monotone in the level of inequality. This implies that policies that decrease inequality without substantially reducing it can leave society more vulnerable to misinformation.

3.2.1 Demonstration of Main Ideas

This section serves as an overview of the technical results in the paper by presenting three examples that demonstrate the complex role of inequality in learning and manipulation. The examples below show that increasing inequality can: $i)$ have divergent effects on different communities, hurting one community and making another better off, or $ii)$ it can hurt the whole society, or $iii)$ it can protect the whole society. This variety of outcomes depends on a myriad of factors like relative community affluence, relative community sizes, and the cost of the manipulation technology.

Below, we go through the details of each of these examples.

Inequality Hurts the Most Marginalized

We consider two communities of equal size and explore the degree of manipulation under different homophily structures. Two of the agents on the first island are knowledgeable, compared to only one of the agents on the other island. Thus, the former island is the *privileged* community and the latter island is the *marginalized* community. This setup is pictured in Figure 3-5. In this example, we vary homophily by setting $p_d = 0.2$ and increasing $p_s > p_d$ (note that as p_s increases, homophily increases). We assume that $\varepsilon = 0$ so that it is costless for the principal to send misinformation (and therefore will send to everyone).

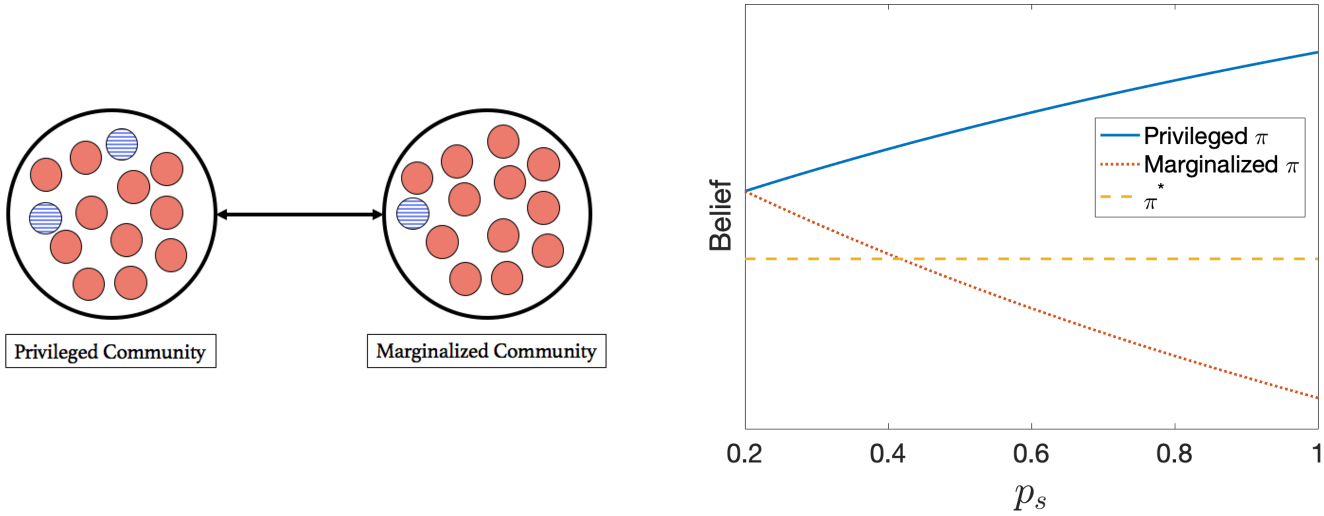


Figure 3-5. On the left is the setup of Section 3.2.1, and on the right are the beliefs of the two communities. As homophily increases (i.e. as p_s increases), the beliefs of the privileged community move towards the truth while the marginalized community's beliefs fall below the belief threshold given by the dashed line, leading to the agents in that community taking the incorrect action.

Figure 3-5 shows the beliefs of the agents in both communities. Recall that manipulation occurs when an agent's belief falls below a certain threshold π^*. We see that as p_s increases, the beliefs of the privileged community move closer to the truth (higher belief) whereas the beliefs of the marginalized community move farther from the truth (lower belief). In addition, there exists a corresponding homophily threshold $\bar{p}_s$ (approximately 0.4 in this example) whereby when $p_s < \bar{p}_s$, there is no manipulation, but when $p_s > \bar{p}_s$, the marginalized community becomes manipulated. An increase in inequality in this example leads to more manipulation in society. Thus, an increase in inequality makes society worse off as the marginalized community becomes susceptible to misinformation.

Inequality Hurts Everyone

We now consider three islands: a small privileged community with an eighth of the population, a small marginalized community with an eighth of the population, and a large community with the remaining three quarters of the population. Assume that there are three knowledgeable agents in the privileged community, one knowledgeable agent in the large community, and no knowledgeable agents in the marginalized community. This setup is depicted in Figure 3-6.

Similar to the previous example, we set $\varepsilon = 0$, fix $p_d = 0.2$, and vary p_s to change the amount

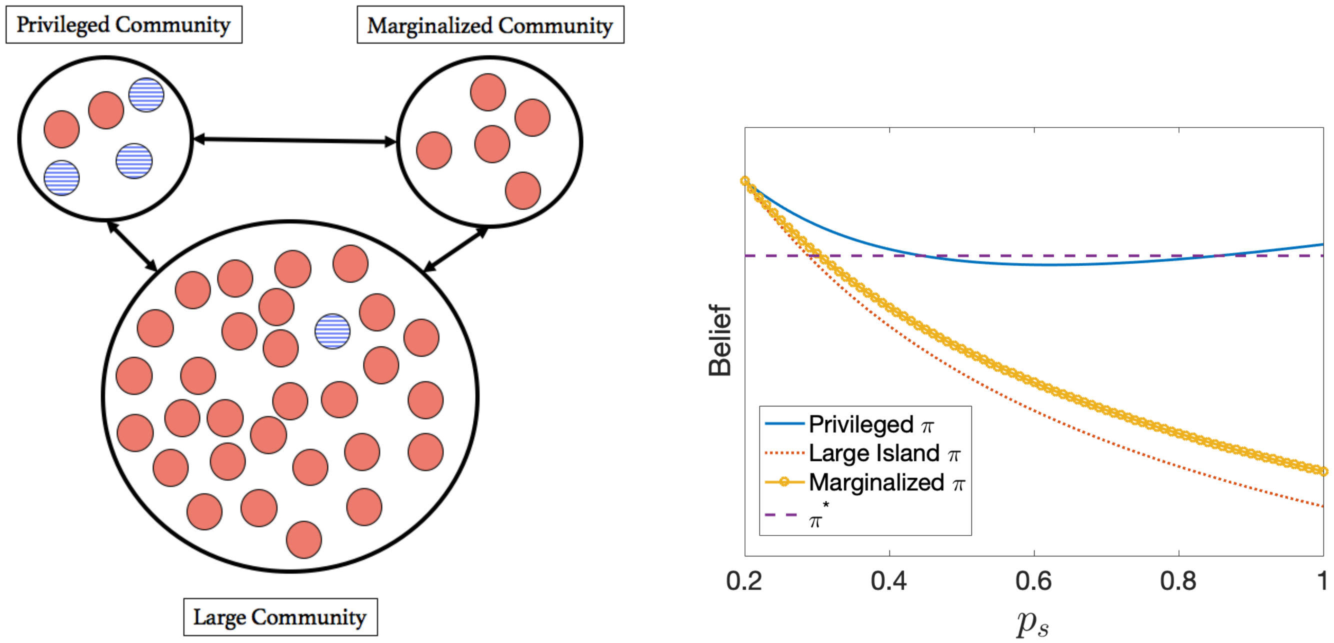


Figure 3-6. On the left is the setup of Section 3.2.1, and on the right are the beliefs of the two communities. As homophily increases (i.e. as p_s increases), the beliefs of all communities move away from the truth and fall below the belief threshold, so that all agents take the incorrect action. Further increase in homophily restores some of the beliefs in the privileged community, but does not bring it back to first-best levels.

of homophily in society. The beliefs of the three different communities are shown in Figure 3-6 as a function of p_s. Given the threshold line π^* in the plot, we see that as homophily increases, the beliefs of *all* agents in the population move farther away from the truth. Once homophily hits $p_s = 0.3$, two communities are manipulated while the privileged community is still immune. As homophily increases further to $p_s = 0.5$, everyone in the network is manipulated. This is true even for the *privileged* community, despite the fact that agents in this community are forming more direct connections with knowledgeable agents who spread truthful information.

Such a phenomenon occurs because the size disparity between communities leads to all of them deriving most of their beliefs from the information spreading in the large community, so when inequality hurts this community, it propagates to those who, on the surface, should be benefiting from it (as in the previous example). Another way of seeing this is that when inequality decreases, knowledgeable agents in the privileged community can have their voices amplified through talking to agents in the large community, who then help spread these beliefs over the network (including back to DeGroot agents in the privileged community).

While intermediate levels of inequality are bad for all agents, Figure 3-6 shows that an increase in homophily beyond a certain point (around $p_s \approx 0.85$) begins to restore the beliefs of the privileged community, but not to the extent of returning these beliefs to first-best levels. This is a consequence of the large community suffering with false beliefs and dragging down the beliefs of those in the privileged community. Thus, once homophily reaches an extreme level, the misinformation rampant on the large island ceases to spread beyond its own community. In that sense, extreme homophily is better than intermediate homophily because at least one of the islands is insulated from the misinformation in the large community.

Inequality Protects Society

We now give an example to show how the spread of misinformation can be shaped by the interplay between the principal's strategy and the inequality structure of society. Consider three communities of the same size. The privileged community has a 3% knowledgeable population, the "average" community has a 1% knowledgeable population, and the marginalized community has no knowledgeable agents. Unlike the previous examples, we assume that $\varepsilon \in (4/5, 1)$, so that it is *not free* for the principal to expend resources in manipulating the beliefs of the agents. This setup is depicted in Figure 3-7.

Figure 3-8a shows the beliefs (of the correct state) when the principal sends signals to *everyone* in the population. Suppose there is no homophily, so that $p_s = p_d = 0.2$, and, as always, there is a belief cutoff π^* for taking the correct action. Then under this strategy, all agents are manipulated, and the cost of sending signals is $\varepsilon < 1$, so this is indeed profitable and the network is susceptible.[21]

As homophily increases, the beliefs of the privileged community move closer to the truth and eventually pass the cutoff, thereby insulating them from the strategy where the principal exerts maximal influence over the entire population. For instance, when $p_s > 0.3$ in Figure 3-8a, the privileged community takes the correct action even though the other two communities do not. However, since $\varepsilon > 4/5 > 2/3$ (the principal is manipulating $2/3$ of the population), a strategy that targets all agents in the population is no longer profitable.

Instead, we investigate whether the principal has a profitable strategy where he incurs less

[21] Note that it is not immediate that every agent will be manipulated in equilibrium, just that the network is susceptible (see Mostagir et al. (2022), Corollary 2). However, it can be shown through a more sophisticated argument that if the principal targets at most 5/6 of the population (regardless of the distribution across islands), then he manipulates no one. Thus, the optimal strategy for the principal is to target sufficiently many agents to guarantee that all islands are manipulated.

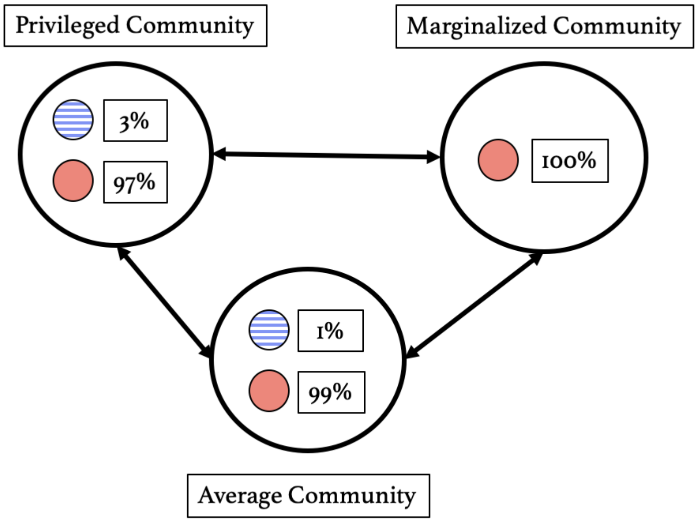
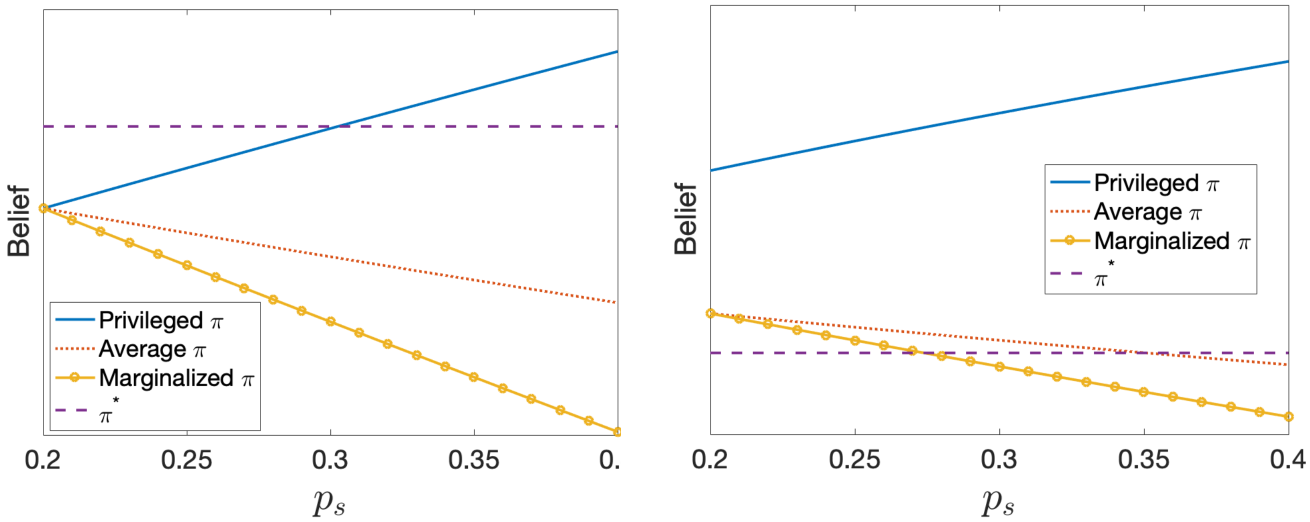


Figure 3-7. Example of Section 3.2.1

cost but still manipulates the average and marginalized communities. It is relatively easy to show that the principal always prefers to decrease his influence on the privileged community, as he cannot manipulate this community anyway, and, because each community has more connections within their own island, exerting influence *within* the average and marginalized communities distorts beliefs the most within these communities. For a profitable strategy then, the principal must abandon sending misinformation to $1 - \frac{2}{3\varepsilon}$ proportion of the population; in particular, given $\varepsilon > 4/5$, he must abandon sending misinformation to 1/6 of the population, or 1/2 of the privileged community. The beliefs of the agents under this strategy are pictured in Figure 3-8b.

Notice though that under this strategy, when $p_s = 0.3$, none of the agents in the average community are manipulated either, while the agents in the marginalized community are. This in turn implies that this strategy is unprofitable, as the principal expends $\frac{5\varepsilon n}{6} > \frac{2}{3}n$ but only receives a benefit of $\frac{1}{3}n$. Instead, the principal must abandon sending misinformation to $1 - \frac{1}{3\varepsilon} > 7/12$ proportion of the population. Once again, it can be shown that the principal is better off targeting the marginalized community directly, before sending misinformation to the other communities. The principal's maximal influence, while still being possibly profitable, comes from sending everyone on the marginalized community misinformation and then either: (i) not sending signals to the privileged community but sending signals to $1/12$ of the average community, or (ii) not sending signals to the average community but sending signals to $1/12$ of the privileged community. The beliefs under both these strategies are pictured in



(a) Beliefs when principal targets all agents

(b) Beliefs when principal sends fewer signals

Figure 3-8. Beliefs of the communities in Section 3.2.1. The left plot shows beliefs when the principal targets everyone. The right plot shows beliefs when the principal sends fewer signals to the privileged island.

Figure 3-9.[22]

As can be seen, under either of these strategies there is not enough misinformation sent to distort even the marginalized island's beliefs. The network is *impervious* to manipulation because of this domino effect: as one community becomes more insulated, its beliefs move closer to the truth and spill over to the next community and the process repeats until all communities are protected and the principal had no profitable strategy. As such, the presence of *some* inequality can end up protecting everyone by discouraging strategic manipulation of beliefs. We call this phenomenon *protection contagion* and explore it further in Section 3.2.3.

3.2.2 Inequality and the Spread of Misinformation

In this section, we focus our attention on the case where the cost of sending misinformation, ε, is close to 0. In such instances, the principal's optimal strategy is trivial: he sends misinformation to everyone in the network. This decouples the belief dynamics and inequality structure from the principal's strategy and allows us to study these dynamics in isolation. The case where $\varepsilon \gg 0$, and when the principal's optimal strategy is non-trivial, is the focus of Section 3.2.3.

[22] Any convex combination of targeting agents in both the average and privileged communities lead to the same conclusion.

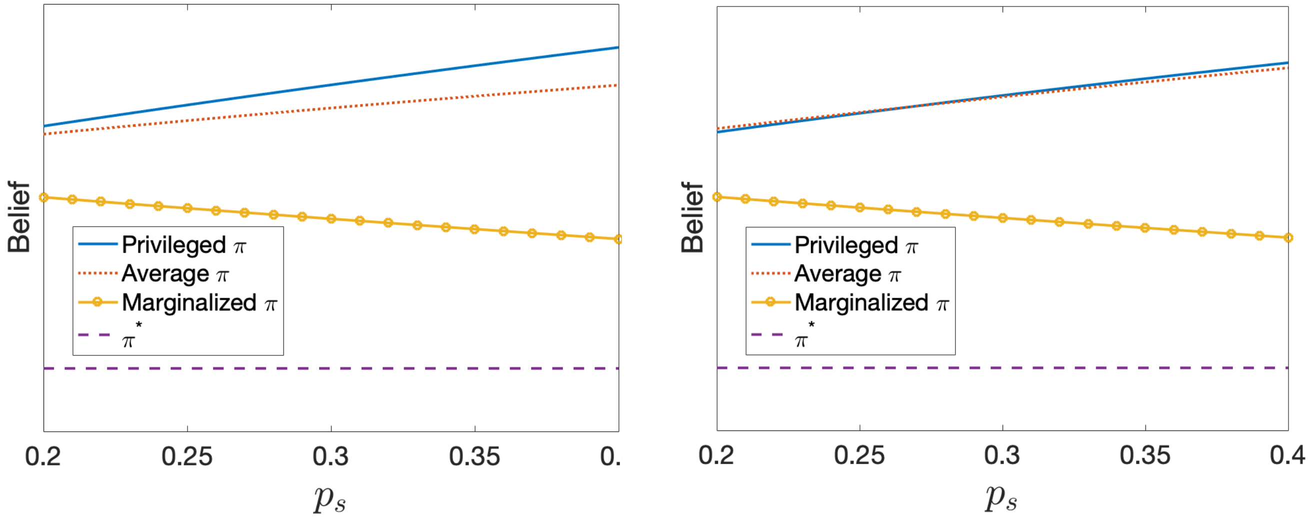


Figure 3-9. Beliefs of the communities in Section 3.2.1 when the principal attempts to manipulate only the marginalized community. On the left (right) are the beliefs when the principal targets a fraction of the average (privileged) community in addition to the marginalized community. Neither strategy is profitable as no one is manipulated.

Inequality and Network Homophily

We assume there are k islands and a society $(p_s, p_d, \mathbf{m})$ is specified by three objects:[23]

1. p_s: the communication within islands (i.e., the within-island link probability).

2. p_d: the communication across islands (i.e., the across-island link probability).

3. $\mathbf{m} \equiv (m_1, \ldots, m_k)$: the vector of knowledgeable agent counts for each island $\ell \in \{1, \ldots, k\}$.

We refer to the pair (p_s, p_d) as the *homophily* of the network; these parameters completely determine the social network structure. We always assume $p_s \geq p_d$. Next we define what it means for one society to exhibit less inequality than another:

Definition 3.2.1. (Inequality) We say that society $(p_s, p_d, \mathbf{m})$ exhibits *less inequality* than society $(p'_s, p'_d, \mathbf{m}')$ if:

(a) There is more communication across islands; namely, $p_d \geq p'_d$;

(b) There is less communication within islands; namely, $p_s \leq p'_s$;

[23] We use "communities" and "islands" interchangeably throughout.

(c) The distribution of knowledgeable agents across groups is more "equally distributed"; formally, $\mathbf{m}'/\mathbf{s}$ is a majorization[24] of $\mathbf{m}/\mathbf{s}$.

with at least one condition strict.

If Society A has less inequality than Society B, this suggests two features. First, any agent in Society A is more likely to talk to agents outside her own island, relative to the higher inequality Society B. This is a direct consequence of less homophily. Second, less inequality also implies less inequality in terms of direct connections to knowledgeable agents: any two agents in Society A are more likely to have a similar number of (weighted) connections to knowledgeable agents as compared to Society B. The most equitable distribution of knowledgeable agents occurs when they are the same constant fraction of the population on every island.

Inequality provides a partial (as opposed to total) ordering on societies. This occurs for the following reasons. First, if we simultaneously increase homophily and more evenly distribute knowledgeable agents, then we create more even access to resources but also restrict how communities share these resources, resulting in an ambiguous inequality comparison. For example, assuming equal island sizes, a society described by $(p_s, p_d, m_1, m_2) = (0.5, 0.5, 3, 0)$ is no more or less equitable than a society described by $(p'_s, p'_d, m'_1, m'_2) = (0.8, 0.2, 2, 1)$. Second, majorization itself defines only a partial order, so it is possible that two knowledgeable agents distributions $\mathbf{m}$ and $\mathbf{m}'$ are not comparable. For instance, assuming equal island sizes, $(m_1, m_2, m_3) = (1, 1, 4)$ is no more or less equitable than $(m'_1, m'_2, m'_3) = (0, 3, 3)$. For this reason, we use the following definition in order to compare inequality structures:

Definition 3.2.2. We say a society has the *most inequality* if there exists no other society with (strictly) more inequality. We say a society has the *least inequality* if there exists no other society with (strictly) less inequality. We say a society has *intermediate inequality* if it is neither a society with the most or least inequality.

Note that a society with the most inequality necessarily has homophily structure $(p_s, p_d) = (1, 0)$ and a society with the least inequality necessarily has $(p_s, p_d) = (0.5, 0.5)$. However, all the analysis that follows is continuous in p_s and p_d, and so holds for (non-empty) open intervals

[24] A majorization $\mathbf{x}'$ of $\mathbf{x}$ satisfies (i) $\sum_{\ell=1}^{k} x_\ell = \sum_{\ell=1}^{k} x'_\ell$ and (ii) $\sum_{\ell=1}^{\ell^*} x_\ell \geq \sum_{\ell=1}^{\ell^*} x'_\ell$ for all $\ell^* \in \{1, \ldots, k\}$, where the components of $\mathbf{x}$ and $\mathbf{x}'$ are sorted in ascending order (see Marshall et al. (2011)). An equivalent condition is whether one can transform $\mathbf{m}'$ into $\mathbf{m}$ via a sequence of "Robin Hood" operations: one can recover $\mathbf{m}$ from $\mathbf{m}'$ via a sequence of transferring knowledgeable agents from islands that have a larger population of such agents to islands with fewer (see Arnold (1987)).

around these homophily parameters as well. Finally, we remind the reader that we use the term *marginalized* to refer to a community that has a smaller proportion of knowledgeable agents compared to a *privileged* community (which has a higher proportion of knowledgeable agents).

Misinformation in Equal-Sized Communities

Recall that when $\varepsilon \approx 0$, the principal's strategy is trivial and she exerts maximal influence on the population, i.e., $\mathbf{x} = \mathbf{1}$. We assume each island has an equal share of the population $s_1 = s_2 = \cdots = s_k = 1/k$. In the same vein as Section 3.2.1, we show that the intuition of "increased inequality is bad for learning" is accurate in the special case where a) islands have equal sizes (as in Golub and Jackson (2012)) and b) the only criterion is whether society as a whole is impervious (i.e. no agent is manipulated) or not, rather than the number of agents manipulated.

Theorem 3.2.1. *If society* $(p_s, p_d, \mathbf{m})$ *is susceptible to manipulation and has less inequality than society* $(p'_s, p'_d, \mathbf{m}')$*, then society* $(p'_s, p'_d, \mathbf{m}')$ *is also susceptible to manipulation.*

In other words, Theorem 3.2.1 states that there is an inequality threshold[25] whereby increasing inequality eventually flips the network from impervious to susceptible. This result corroborates the evidence that inequality hurts learning; in particular, inequality always negatively affects learning in the most marginalized communities. However, Theorem 3.2.1 does not claim that total manipulation —the *number* of manipulated agents— is monotone in the degree of inequality. In particular, once the network becomes susceptible, it may be possible that increasing inequality leads to a reduction in the extent of manipulation, though it does not return the network to its first-best state of imperviousness. This property holds generally:

Theorem 3.2.2. *For any society with* $k \geq 3$ *islands of equal size and* m *total knowledgeable agents :*

(i) For a given b, m*, if there is an impervious network for some inequality structure, the network with the least inequality is impervious;*

[25] Because there is only a partial ordering of societies, this threshold holds two of three inequality parameters constant while changing the third one. For example, if the knowledgeable agents distribution is the parameter being changed, then one can apply the threshold for any partially ordered sequence of distributions.

(ii) For all b, m, there always exists a network with intermediate inequality that has (weakly) more manipulation than some network with more inequality.

(iii) There exist values for b, m such that the network of (ii) with intermediate inequality has strictly *more manipulation than some network with more inequality.*

Theorem 3.2.2 states that an "intermediate" amount of inequality is worse than an extreme amount of inequality, which in turn is worse than no inequality at all. While removing all inequality improves learning, simply reducing inequality in an extremely homophilous society can actually lead to worse learning and manipulation outcomes.

Underlying the previous result is the fact that social connections have both positive and negative externalities. On one hand, they serve as a transmission mechanism for spreading the (correct) beliefs of knowledgeable agents. However, they also allow the principal to spread misinformation in a more effective way, by using social forces to manipulate other agents as well. When homophily is extreme, the principal cannot use one community to influence another. These missing connections can prevent the principal from manipulating certain communities, who had previously derived their beliefs from more marginalized communities when homophily was not too extreme. On the other hand, when homophily is quite weak, access to knowledgeable agents is relatively similar across islands, which allows them to communicate truth most effectively. It is the intermediate homophily case that often acts as a perfect breeding ground for manipulating beliefs.

This result provides a sleek connection to models of contagion in financial networks (see Acemoglu et al. (2015), Babus (2016), Kanak (2017), for example). Similar to the degree of homophily in our setting, in these models, connections both serve to reduce and exacerbate the propagation of negative forces. On one hand, when a bank's linked institutions are in distress, the bank finds itself less well-capitalized and more likely to default. However, when a bank faces an idiosyncratic or temporary problem, it can rely on neighboring (safe) institutions to protect it from insolvency. Hence, the stability of a financial network can be subtle, and the effect of increased interconnectivity is typically ambiguous, just as with social learning in the presence of homophily and inequality.

3.2.3 Strategic Influence and Inequality

The results in Section 3.2.2 are obtained under the assumption that the cost of sending misinformation is negligible, and therefore the principal targets every agent in the population. This $\varepsilon = 0$ case enabled us to measure how misinformation propagates as a function of the inequality structure in society, without introducing strategic considerations on the part of the principal.

We now relax this by assuming $\varepsilon \gg 0$, and for simplicity also assume that all communities are the same size. The latter assumption allows us to isolate the effects of the principal's strategy from the population size effects identified in Sections 3.2.1 and 3.2.4. The $\varepsilon \gg 0$ assumption requires a strategic choice by the principal of who to target, and is of critical importance in understanding the spread of misinformation in the presence of a strategic actor.

For the interested reader, Appendix B.2.2 expands on Section 3.2.1 and provides a detailed walkthrough of an example where ε varies from small to large in a society with two communities, and shows that *some* inequality can protect the entire society by initially protecting the privileged community. Theorem 3.2.3 extends this to an arbitrary number of communities by showing that such a network where intermediate inequality is best for society always exists when the principal faces non-negligible signaling costs. This contrasts with Theorem 3.2.2, where intermediate inequality is not only never optimal, but is always (weakly) worst for society when signaling costs are low. Finally, we conclude with some numerical experiments that show how manipulation changes as a function of simultaneously varying the investment cost and the inequality structure under the principal's optimal strategy.

Protection Contagion: The Case for *Some* Inequality

Recall from Section 3.2.1 that when there was no inequality, the principal had a profitable strategy to target and manipulate everyone. With some inequality, however, the principal was unable to manipulate one of the more privileged communities, which in turn made his strategy too expensive. To maintain a profitable strategy, the principal had to reduce his direct influence on that community in order to save costs, while still trying to retain the same extent of (indirect) overall influence on the other communities. However, this reduction made the principal unable to manipulate the next privileged community, which similarly led to him reducing his direct influence in that community, and so on. We refer to this cascade effect as *protection contagion.*

This effect is not an artifact of Section 3.2.1, or the example presented in the previous section. In fact, when the principal has intermediate costs for sending misinformation, protection contagion can sometimes lead to a complete unraveling of his influence when there is some inequality in the network. This is summarized in the next result.

Theorem 3.2.3. *Suppose there are $k \geq 2$ islands of equal size. There exists $b^* < b^{**}$, $\varepsilon^* < \varepsilon^{**}$, such that if $b \in (b^*, b^{**})$ and $\varepsilon \in (\varepsilon^*, \varepsilon^{**})$, there exists a network with intermediate inequality that is impervious, despite every network with the most inequality being susceptible, and every network with the least inequality admitting strictly more manipulation than networks with the most inequality.*

Theorem 3.2.3 describes a range where intermediate inequality is best for protecting society from the spread of misinformation. Suppose we order the communities based on their privilege, i.e. the proportion of knowledgeable agents in the population, and protect the most privileged community from manipulation. This protection forces the principal to decrease his effort in this community to try and maintain a profitable strategy. By doing so, the beliefs in that community move closer towards the truth, and because there is still *some* communication across communities, this provides a positive externality to the rest of the network. The principal then is unable to manipulate the next privileged community, and so stops targeting that community as well, leading to a recursive process that repeats for all communities, and the principal cannot target anyone while retaining a positive payoff. However, if inequality becomes extreme, this contagion effect fails to take place: the positive spillovers from protecting one community are minimal in the face of gross inequality. Extreme homophily leads to little communication across communities and so protecting one community still leaves the rest exposed to misinformation.

Note the connection between Theorem 3.2.3, when $\varepsilon \gg 0$, and Theorem 3.2.2(a), when $\varepsilon \approx 0$. Theorem 3.2.2(a) states that if some inequality model is impervious, then the least inequality attains imperviousness. This is not the case for $\varepsilon \gg 0$; in particular, Theorem 3.2.3 states that it may be possible for an intermediate inequality model to be the only model that attains imperviousness.

Numerical Simulations

We provide results from two numerical simulations that illustrate the non-monotonic behavior from the previous section on the broader parameter space. Recall that we can increase the level of inequality by increasing homophily or by having a more uneven distribution of knowledgeable agents between islands of equal size. We simulate both of these scenarios. In the first simulation, we vary the extent of homophily through varying p_d (while holding p_s fixed). The second simulation varies the distribution of knowledgeable agents across the islands. In both cases, we simultaneously vary the cost ε that the principal faces.

Homophily. We fix $p_s = 0.8$ and take $b = 0$, so that an agent takes an action based on the state she believes is most likely. There is a total population of 1000 agents split equally across two islands; one island has 80 knowledgeable agents and the other has the remaining 20.

The left heat map in Figure 3-10 shows the results of this simulation. In the range of $\varepsilon \in (1.1, 1.7)$, we notice the non-monotonicity described in Theorem 3.2.3 as we increase p_d (i.e., decrease homophily/inequality). For small values of p_d (large homophily), half the agents are manipulated. As we decrease homophily through increasing p_d, we transition to a region where the network is impervious. Finally, as homophily decreases further, we end up in a region where *all* agents in the network are manipulated. This is the same effect seen in the example of Appendix B.2.2.

Distribution of Knowledgeable Agents. We fix $(p_s, p_d) = (0.5, 0.2)$ and take $b = 0$. There is a total population of 1000 agents, split over two islands of equal size, and we vary the number of knowledgeable agents, m_1, on the first island from 0 to 100 (with the other island containing the remainder, $m_2 = 100 - m_1$) .

The results are shown in the right heat map in Figure 3-10. Inequality between islands is most severe when $m_1 = 0$ or $m_1 = 100$, with the least inequality at $m_1 = 50$. In the range $\varepsilon \in (0.9, 1.7)$, we see that the network is impervious provided there is sufficient inequality in the distribution of knowledgeable agents; otherwise, all agents are manipulated. This inequality protects one island from manipulation and, through protection contagion, prevents the principal from having any profitable strategy.

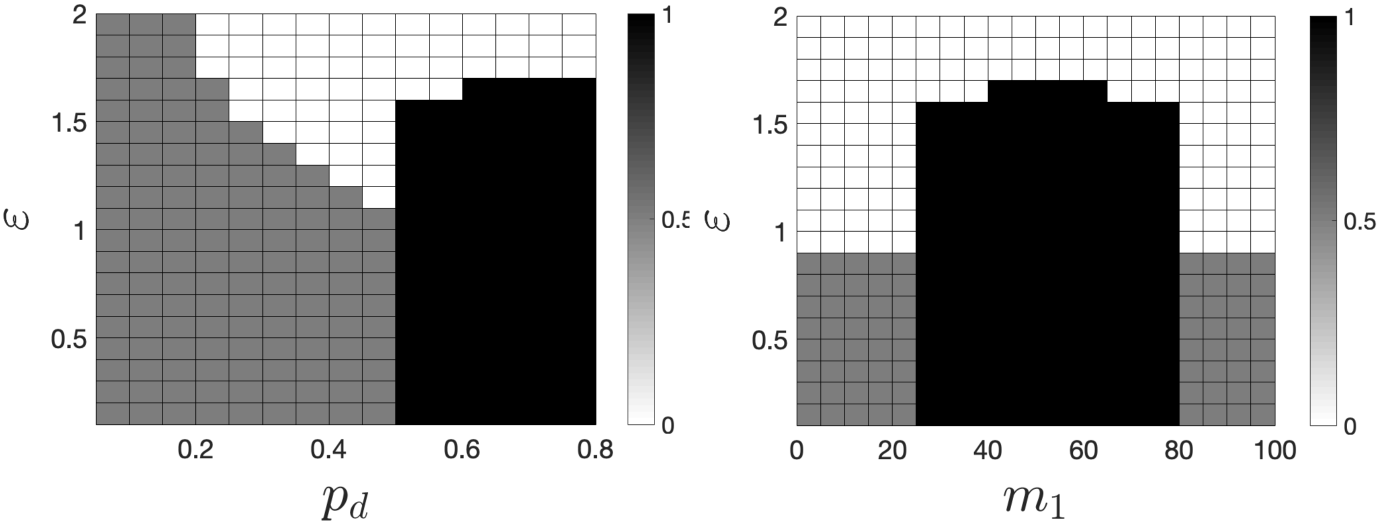


Figure 3-10. Heat maps showing fraction of agents manipulated as a function of the signaling cost ε and p_d (left figure) or m_1 (right figure). Note that increasing p_d implies decreasing homophily/inequality. Light blocks indicate no manipulation, while gray (dark) blocks indicate half (all) the population is manipulated.

3.2.4 Different Community Sizes and Strong Inequality

Up until now, we have focused on communities that are roughly equal in population and whereby agents in those communities associate more with those in their own community, but do not differentiate their social interactions amongst other communities. In this section, we consider the complexities of having both (i) large or small communities and (ii) inequality structures that are "strong," in the sense that there are significant barriers to communication between certain communities. For the latter, we consider an inequality structure where agents further differentiate their social interactions by only affiliating with groups who have characteristics that are *close* to those of their own community. We call this *strong inequality*. We consider how this type of strong inequality affects society at large relative to the weaker notion of inequality with a flat hierarchy, as we have studied in the previous sections.

Misinformation with Different Community Sizes

We now consider the case when communities are not the same size. We begin with the following definition:

Definition 3.2.3. We say that an island ℓ is *least privileged* if (i) its belief of the correct state is the least of any island (i.e., $\pi_\ell \leq \pi_{\ell'}$ for all ℓ') and (ii) it has the least knowledgeable agent percentage of the population (i.e., $m_\ell/s_\ell \leq m_{\ell'}/s_{\ell'}$ for all ℓ').

Observe that condition (i) is also equivalent to island ℓ having the largest DeGroot centrality. Note that with islands of the same size, condition (i) holds if and only if condition (ii) holds for island ℓ, so is redundant. However, with islands of different sizes, because influence is asymmetrical across islands, neither condition implies the other.

As we saw in Section 3.2.1, when communities have different population sizes, the results of the previous section need not hold. When there is a large community, it is possible that additional inequality can hurt *the entire society*. Because most communities draw their beliefs from the belief of the "masses," the effect of inequality on the masses determines how society as a whole is affected by inequality:

Theorem 3.2.4. *Suppose there are k islands of unequal sizes. Assume the largest island, island 1, is the least privileged. For almost all b, there exists size threshold $\bar{s}$ such that if $s_1 > \bar{s}$, the number of manipulated islands is monotonically increasing in inequality, provided that island 1 remains the least privileged.*

Theorem 3.2.4 states that if we are to decrease inequality with a large least privileged island, manipulation can only decrease. Put more simply, if the masses are the least privileged, then decreasing inequality helps everyone, including very privileged communities. This is because these communities still form a sizable number of connection with the large island, just by virtue of the size disparity, and hence draw a large part of their beliefs from there. Indeed, as inequality decreases, knowledgeable agents in privileged communities can have their voices amplified through talking to agents in the large community, who then spread these beliefs over the network (including back to DeGroot agents in the privileged communities). The flip side of this is that if the masses are the least privileged, increasing inequality helps no one: in fact, moving resources from the masses to the privileged communities ends up making both the masses and the privileged communities worse off. Theorem 3.2.4 thus establishes that inequality benefits society as a whole (in a Pareto sense) if it benefits the large community that wields heavy influence. Likewise, even the privileged islands should want to move their resources to reduce inequality.

Note the assumption that island 1 is the least privileged (and remains so after decreasing inequality) cannot be dispensed with. If island 1 is simply underprivileged, non-monotone comparative statics might exist following an increase in inequality. The intuition is as follows. While this inequality hurts island 1's access to more privileged communities' resources, it also exposes island 1 less to communities which are less affluent than itself and may have more misinformed beliefs. This effect does not exist, of course, when island 1 starts off as the least well-off community.

Weak Inequality vs Strong Inequality

We introduce a different stochastic-block model where are communities ordered by similarity, with agents in neighboring communities more likely to be linked than agents in communities that are farther apart. This model captures the more hierarchical structure that is sometimes observed in society. While this is a natural homophily model, we are not aware of any literature that studies it compared to the much stronger focus on weakly-assortative networks that we studied up to this point in the paper . In this model, each community ℓ has a vector of qualities, $\Lambda_\ell \in \mathbb{R}^L$. Qualities can capture different variables like education, profession, income, etc. Communities are sorted according to their similarity, with the distance metric between communities ℓ and ℓ' given by $d(\ell, \ell') = ||\Lambda_\ell - \Lambda_{\ell'}||_2$. For simplicity, we assume that $L = 1$ (the quality vector is one-dimensional) and thus communities are (strongly) ordered by their Λ_ℓ on a *line topology*.

In both the weak and strong inequality models, for any two agents on the same island, there is a link probability p_s. In the weak inequality model, there is also a link probability $p_d < p_s$ for any two agents on different islands. However, in the strong inequality model, agents do not form links with agents on islands outside of their neighboring islands. Agents in community ℓ are linked to agents in community $\ell - 1$ or $\ell + 1$ with probability p_d, whereas agents in "farther" communities are linked with probability 0,[26] with the exception of island 1 and island k, which are linked to island 2 and island $k - 1$ only, respectively.

Section 3.2.2 and Section 3.2.3 documented the effects of weak inequality on manipulation. To make the comparison between strong and weak inequality most transparent, we consider *worst-case* inequality for weak inequality, i.e., the inequality structure that makes the principal most easily able to manipulate. Toward this end, the following result establishes a condition

[26] We can equivalently assume that these link probabilities are positive but decay sufficiently quickly, such as on the order of $\exp(-||\Lambda_\ell - \Lambda_{\ell'}||_2)$. For simplicity of exposition and illustration of the effects of our strong assortative property, we simply set the link probabilities to 0.

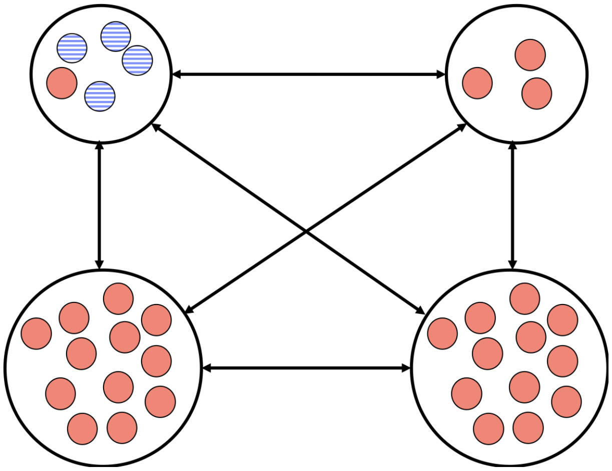


Figure 3-11. An illustration of Proposition 3.2.1: under weak homophily, a linear number of knowledgeable agents is enough to prevent manipulation anywhere in the network. This could include stacking them all on one island.

on the *total number of knowledgeable agents* in the weak homophily model needed for imperviousness, *independent* of their placement across communities:

Proposition 3.2.1. *For any (p_s, p_d), there exists $\bar{\theta}$ such that if $\theta < \bar{\theta}$, there exists a constant $c < 1$ such that if there are $m = cn$ knowledgeable agents anywhere, then* any *weak inequality model (regardless of the number of communities k) is impervious. Moreover, c is increasing in p_s and decreasing in p_d.*

In other words, there exists a threshold c whereby if a proportion c of the population is knowledgeable, the principal will be unable to manipulate anyone, regardless of the depth of weak inequality present. This includes the most extreme inequality configuration where all the knowledgeable agents are on one island, and the rest of the $k - 1$ islands are all DeGroot (for any k). A visual depiction of Proposition 3.2.1 is given in Figure 3-11. The assumption that θ is not too large ensures that agents use social learning as a primary means of learning; clearly when θ is too close to 1, the presence of knowledgeable agents is irrelevant because agents place too much weight on their own (manipulated) news.

Proposition 3.2.1 also sheds some light on whether homophily helps or hurts the worst-case lower bound. Because m is increasing in p_s and decreasing in p_d, we see the number of knowledgeable agents needed to apply Proposition 3.2.1 increases as we increase inequality through the network homophily structure. This result reinforces the general idea that increasing inequality makes it more challenging for society to avoid manipulation, despite the exceptions presented earlier. The intuition is clear: as homophily becomes more severe, configurations like that of Figure 3-11 do little to help communities with few to no knowledgeable agents.

For illustration, we assume that the first community has m knowledgeable agents and all other communities consist of DeGroot agents. At the end of this section, we discuss the robustness of the result to other configurations. There are k communities which may or may not be the same size and we fix (p_s, p_d). In the strong inequality model we obtain a much different result from Proposition 3.2.1:

Proposition 3.2.2. *For any $\theta > 0$ and $c < 1$, there exist $\bar{k}$ (independent of k) and $\varepsilon > 0$ where all communities except $\bar{k}$ are manipulated, even with cn knowledgeable agents.*

Proposition 3.2.2 shows the stark difference between weak and strong inequality. First, with weak inequality, we can always find a proportion c such that cn knowledgeable agents will make the network impervious, even with rampant (weak) inequality. On the other hand, we can never find such a proportion c in the strong inequality model: no constant fraction guarantees society is safe from manipulation because the influence of knowledgeable agents is too diluted under strong homophily, as seen in Figure 3-12. Second, the strong inequality network is not only susceptible, but manipulation is actually *ubiquitous* in society. Note that $\bar{k}$ does not depend on k, so when there are several communities, only a vanishing fraction of them will not be manipulated. Except for a very small set of communities who happen to have close ties to knowledgeable agents, almost all communities will be negatively impacted by the existence of strong inequality.

The intuition for the result is as follows. Notice that with strong inequality, as in Figure 3-12, agents receiving misinformation communicate their beliefs both forwards and backwards, which leads to more propagation of misinformed beliefs. This creates a strong *echo chamber effect*, where the influence from misinformation, as reflected in the agents' beliefs, gets inflated because they fail to recognize their own influence on their own neighboring islands' beliefs. For agents who are not in communities extremely close to the knowledgeable agents' community, this echo chamber is strong enough to completely mask any influence the knowledgeable agents might have in spreading accurate beliefs. Contrast this with weak inequality in Figure 3-11, where every community has some direct interaction with knowledgeable agents, even if those agents do not reside on that community. This not only provides a direct positive influence on everyone's beliefs, but also prevents these echo chambers from wielding too much power, precisely because other communities are also directly interacting with knowledgeable agents.

Robustness. Finally, we consider how robust these strong inequality results are to the initial setup. Suppose instead of stacking all of the knowledgeable agents on the first island, we instead redistribute them in a way that dampens these echo chamber effects. Would this mitigate the effects of Proposition 3.2.2? An affirmative answer to this question requires this redistribution to be significant. From Proposition 3.2.2, it is easy to see that any island with a knowledgeable agent cannot protect more than a constant number of communities on either side of it. Therefore, the number of knowledgeable agents would need to be dispersed very evenly across all communities to have any hope of preventing manipulation. For example, simply moving the knowledgeable agents to a more central community or distributing them across a couple of islands throughout would have no significant effect, and the conclusion of Proposition 3.2.2 remains intact. Thus, while agents can be protected in the strong homophily model, the requirements on the knowledgeable agents distribution are much stricter: nearly every island has to have some knowledgeable agents of its own, which requires drastically less inequality.

Second, Mostagir et al. (2022) show that higher *density*, while not a perfect measure, is often related to lower manipulation (for instance, see Theorem 4 in Mostagir et al. (2022)). It is clear that the average degree with strong inequality will be lower than that of weak inequality, so a natural question is to wonder whether this difference in density is what drives the difference in manipulation we observe between the two models. For concreteness, assume we have k communities of the same size in both the strong (with p_s, p_d) and weak inequality (with p'_s, p'_d) models. In the strong inequality model, we take $p_s = \alpha p'_s$ and $p_d = \alpha p'_d$, where $\alpha = \frac{p_s+(k-1)p_d}{p_s+2p_d}$.[27] This equalizes the average degree (i.e., connections) of the strong and weak inequality models, but has no effect on any of the beliefs of the agents (or on their DeGroot centralities).[28] Therefore, we see that the differences in density alone cannot explain the differences seen across the two models.

[27] For this, we have to naturally assume p'_s and p'_d are not too large so that $p_d < p_s < 1$ and this is possible. Otherwise, there is no way to equalize the average degrees of the two models.

[28] Technically, this equalizes the average degree for only the islands in the "middle" of the line, but not those on the ends. However, assuming there are a large number of communities, this difference will be negligible.

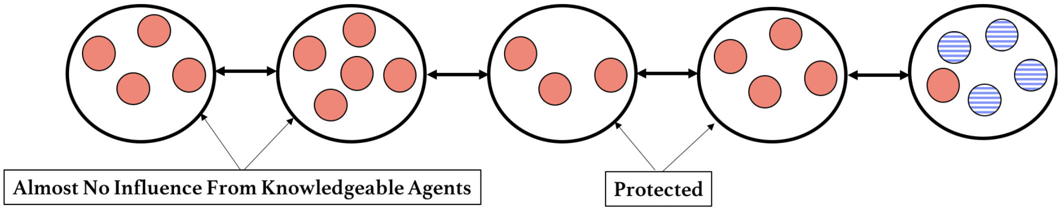


Figure 3-12. An illustration of Proposition 3.2.2: even with many knowledgeable agents, strong homophily allows the principal to manipulate plenty of agents in the network. In the figure above, communities which are not “close enough” to the knowledgeable agents will not be very influenced by their beliefs, so will be manipulated.

3.2.5 Optimal Interventions

We now discuss the role that a social planner has in combating misinformation. We consider two possible interventions: educational interventions and homophily interventions. In the former, we assume the planner may improve the sophistication type of a subset of agents, perhaps through targeted education. In the latter, the planner may decrease the extent of homophily through efforts to integrate communities (i.e. by increasing p_d). The social planner wants to enact a policy that protects as many agents as possible from manipulation.

We say a policy is *optimal* if it minimizes the number of manipulated agents. Similarly, we say some policy X *dominates* another policy Y if all agents' beliefs of the correct state are higher under X than under Y. While an optimal policy is never dominated, there may be non-optimal policies that are not dominated, and thus lie at the Pareto frontier of effective interventions.

We can write the beliefs of the agents, $\boldsymbol{\pi}$, as:

$$\boldsymbol{\pi}(p_s, p_d, \mathbf{m}, \mathbf{x}) = \left(\frac{\mathbf{I}}{1-\theta} - \boldsymbol{B}(p_s, p_d, \mathbf{m}, \mathbf{x}) \right)^{-1} \boldsymbol{a}$$

where $\boldsymbol{B}$ is a function of (i) the homophily structure (p_s, p_d), (ii) the distribution $\mathbf{m}$ of knowledgeable agents across islands $\mathbf{m}$, and (iii) the principal's strategy $\mathbf{x}$. Recall that the belief (of the correct state) threshold is given by $\frac{1+b}{2}$ and the principal wants to maximize the number of agents whose beliefs fall below this threshold, less the total cost of manipulation, so as before, the principal solves:

$$\mathbf{x}^*(p_s, p_d, \mathbf{m}) = \arg\max_{\mathbf{x}} \sum_{i=1}^{n} \left(\mathbb{1}_{\boldsymbol{\pi}_i(p_s, p_d, \mathbf{m}, \mathbf{x}) < (1+b)/2} - \varepsilon x_i \right)$$

Then the planner solves the min-max optimization problem:

$$\min \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{\pi}_i(p_s,p_d,\mathbf{m},\mathbf{x}^*(p_s,p_d,\mathbf{m}))<(1+b)/2}$$

The combinatorial nature of these problems preclude a general solution. We derive the optimal policies for some special cases and show the nuances of optimal policies via simulation. For a number of cases, we prove the optimal policy attempts to minimize inequality. However, this is not always true: if the planner cannot completely eradicate inequality, then sometimes measures that only slightly reduce it can be counterproductive.

Educational Interventions

We consider the possibility of endowing some agents in the population with verifiable knowledge about the true state. We assume that this process is costly and that the planner's budget constraint is of the form $\sum_{i=1}^{n} \mathbb{1}_{\text{type}(i)=K} \leq M$, where K designates a knowledgeable agent and M is an integer. Note that the planner will always use the entire budget in an optimal policy.

Intervention with Large Budget and Cheap Signals. First, we derive the optimal policy when the planner's budget is sufficiently large and the principal's cost of sending signals is nearly free:

Corollary 3.2.1. *Suppose that the budget M is large enough so that it is possible for the planner to make the network impervious when ε is small. Then if all islands are the same size, the optimal policy is to minimize inequality.*

This result is in-line with the conclusion of Theorem 3.2.1, which argues that when imperviousness is possible, the least inequality is always first-best. As a special case, if M is big enough to make the knowledgeable agents equal on every island, then this is the optimal policy.

Note the assumption that M is big enough that imperviousness is attainable for small ε is necessary. First, if imperviousness is attainable only for a given $\varepsilon \gg 0$, then it is possible that a configuration with some inequality may be optimal, as we show next. This is a direct consequence of Theorem 3.2.3. Second, if M is small and so imperviousness is impossible, then an optimal policy may involve creating some inequality to protect at least a fraction of the population. For instance, placing knowledgeable agents equally may lead to every island being manipulated, whereas stacking them all on one island would protect at least this island.

Intervention with Costly Signals. We simulate the optimal policy for the planner when the principal's signals are costly. Consider Figure 3-13 with budget $M = 4$, $n = 100$, two islands of equal size, and homophily structure $(p_s, p_d) = (0.8, 0.2)$. We investigate whether the optimal policy can make the network impervious as a function of ε.

As ε ranges from 0 to 0.5 (small ε region), there is no distribution of knowledgeable agents that admits imperviousness. Corollary 3.2.1 allows us to quickly check this by distributing these agents evenly across the two islands and checking if manipulation exists, which it does. As ε becomes slightly higher than 0.5, the most inequitable distribution (4 knowledgeable agents on one island and 0 on the other) or the most equitable distribution $(m_1, m_2) = (2, 2)$ leads to manipulation. On the other hand, an unequal distribution of $(m_1, m_2) = (3, 1)$ or $(m_1, m_2) = (1, 3)$ leads to imperviousness. When ε continues to increase, *only the even distribution* $(m_1, m_2) = (2, 2)$ *makes society susceptible*, i.e., $(0, 4)$ and $(4, 0)$ are also impervious, and splitting the knowledgeable agents equally across both islands is the worst distribution for society. Of course, eventually, all distributions are impervious when the cost becomes too prohibitive for the principal to have a profitable strategy. In summary, for the planner, the most equitable distribution $(m_1, m_2) = (2, 2)$ is weakly dominated by every other distribution over the *entire* cost range of ε.

Therefore, in the case of costly signals for the principal, the planner may want to introduce some inequality in the knowledgeable agents distribution. This is precisely to generate the protection contagion effect documented in Theorem 3.2.3.

Intervention with One Large Island. We next consider a setting where there is one large island and many small islands; without loss, let the large island be island 1:

Corollary 3.2.2. *When ε is small, there exists $\bar{s}$ such that if $s_1 > \bar{s}$, any policy that makes island 1 the least privileged is dominated by a policy with more knowledgeable agents on this island, provided that such a policy does not already use the entire budget on island 1.*

Corollary 3.2.2 complements Theorem 3.2.4: when there is a single large island, a policy which subjects the masses (on the large island) to inequality is not only sub-optimal, it actually hurts *all* agents in society. In particular, if we start with a configuration of many small

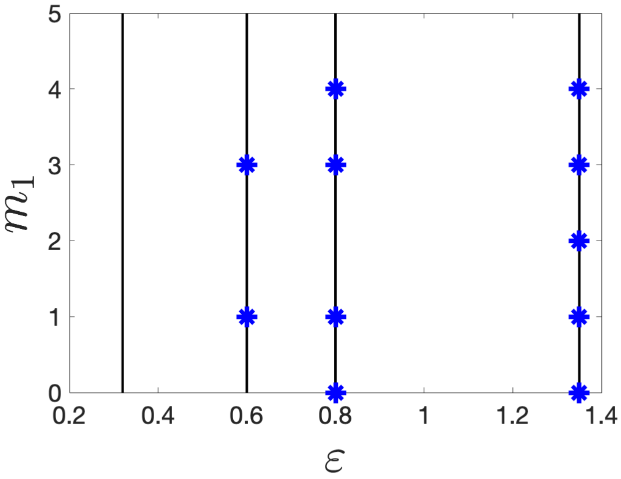


Figure 3-13. Policies that obtain imperviousness with budget $M = 4$. The number of knowledgeable agents on the first island is m_1 and ε is the signaling cost of the principal. Every highlighted point is a knowledgeable agent placement that leads to imperviousness, e.g. $m_1 = 1$ and $m_2 = 4 - 1 = 3$ when $\varepsilon = 0.6$.

privileged communities and one large under-privileged community, then even the privileged communities are negatively impacted by taking resources (knowledgeable agents) from the under-privileged community. Therefore, the social planner *and* those agents in privileged communities should (rationally) support expending more resources on the least privileged community.

Homophily Interventions

We now consider a fixed homophily model with parameters (p_s, p_d^o) and knowledgeable agents distribution $\mathbf{m}$. We assume the social planner pays a positive, convex cost $\phi(p_d - p_d^o)$ with $\phi(0) = 0$ to increase (or decrease) connections between islands. As before, we assume the planner has a budget to spend; that is, the planner must satisfy $\phi(p_d - p_d^o) \leq Budget$.

For the remainder of this section, we focus on the case where the principal's signaling cost ε is small. Similar conclusions to the previous section apply when $\varepsilon \gg 0$. Our first result shows that an equally-distributed knowledgeable agents policy eliminates the need for a homophily intervention:

Proposition 3.2.3. *If knowledgeable agents are equally distributed (i.e., $m_\ell = Ms_\ell$ for all ℓ), then $p_d = p_d^o$ is an optimal policy.*

When knowledgeable agents are distributed proportional to the islands' populations, the beliefs of all agents in society are the same regardless of the homophily parameters. Therefore, no additional intervention is necessary because access to knowledgeable agents is perfectly equitable.

Large Budget. We first focus on the case where the planner's homophily budget is fairly large. Based on our observations in Theorem 3.2.1 and Theorem 3.2.2 we have:

Proposition 3.2.4. *Suppose the budget is greater than $\phi(p_s - p_d^o)$ and large enough to make the network impervious. Then if all islands have the same size, $p_d = p_s$ is the optimal policy when ε is sufficiently small.*

When homophily can be fully corrected, some intervention is desirable. By Theorem 3.2.2, we know that if the budget is big enough for the planner to implement $p_d = p_s$, this obtains the first-best outcome. Thus, the optimal policy is always to completely eliminate any homophily that exists in the network.

Small Budget. When the planner's budget is limited, implementing $p_d = p_s$ may not be feasible. In this case, it may not be optimal to simply reduce inequality by minimizing $(p_s - p_d)$; as we saw in Theorem 3.2.2, often some inequality can be worse than extreme inequality. In other words, a planner who simply helps decrease inequality without eradicating it completely can do unintended harm. In fact, unlike educational interventions, the planner may not want to use the entire budget.

We simulate the optimal p_d with three islands of equal size, $n = 999$, which have 100, 60, and 10 knowledgeable agents, respectively. We assume the cost function $\phi(\alpha) = 10\alpha^2$, $p_d^o = 0$, and $p_s = 0.8$. Note the principal can set $p_d \leq \sqrt{Budget/10} < 0.8$ provided that the budget is at most 5. As in Proposition 3.2.3, we assume the cost of the principal's signaling technology is small, i.e., $\varepsilon \approx 0$.

In Figure 3-14, we show both the optimal homophily (p_d) choice of the planner and the maximum p_d attainable by the budget. We see that when the budget is small, the planner prefers to leave relatively extreme homophily in society as opposed to making the more substantial correction allowed by the budget. Once the budget exceeds a threshold (around 1.3 in the figure), however, the planner uses all of it up to remove as much homophily from the network as possible. Thus, a planner tasked with stopping misinformation through reducing inequality

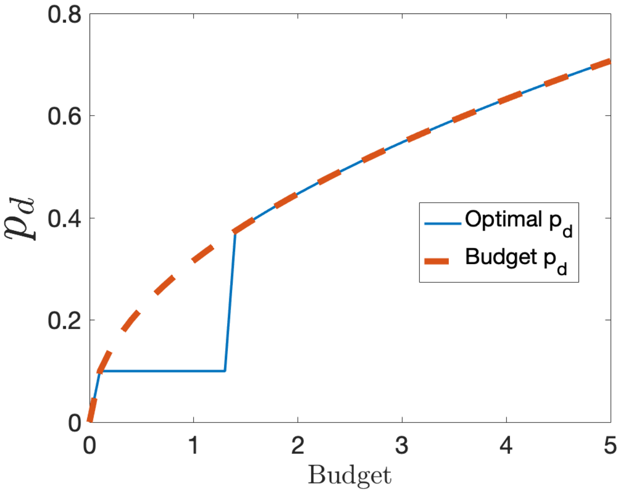


Figure 3-14. Optimal (Simulated) p_d with limited budget. The dotted curve is the level of homophily achieved by spending the entire budget available, whereas the solid curve is the optimal level of homophily given that budget. A planner should always ask for a minimum budget (around 1.3 in the figure) allocation that makes these two curves coincide.

should always ask for a budget allocation that is at least equal to that threshold, in order to guarantee that this reduction will indeed achieve the desired effect and be beneficial to society.

3.2.6 Conclusion

This paper analyzes the role of inequality in social learning when the information that agents receive is a mixture of organic news and news originating from a strategic actor. This setup resembles many scenarios where an information provider may have their own agenda and exerts costly effort to influence agents to take certain actions. Inequality in society results from the distribution of knowledgeable agents who know the true state of the world. Privileged communities have a higher proportion of these agents, and the homophily structure of the network determines access to these agents across communities.

We show that the role that inequality plays in the spread of misinformation is shaped by relative community privilege, relative community sizes, and the cost of the principal's signaling technology. This leads to a range of outcomes depending on how these factors interact. For example, when the privileged communities are small in size compared to the population at large, as is often the case, then an increase in inequality not only makes the large population worse off, but it also makes the privileged communities themselves more prone to misinformation, thus it is in the privileged communities' best interest to encourage allocating resources to the larger community. Generally, the spread of misinformation in not monotone in the level of inequality in the network. Even more so, intermediate levels of inequality can be worst for society when the signaling costs of the principal is low, but can be best for society when the principal's costs are high.

In a similar vein, policies that counteract manipulation depend on the principal's signaling costs as well as the social planner's budget. When signaling costs are low, so that the principal can target everyone, then the planner's optimal policies with a large budget involve eradicating inequality through an equitable placement of knowledgeable agents and/or removing homophily from the network. When signaling costs are no longer trivial, the planner needs to be more careful, as inequality extremes might be worse for society than intermediate inequality regimes. This is a consequence of a protection contagion phenomenon that precludes the principal from having a profitable manipulation strategy. Generally, the complexity of computing these optimal strategies provides a wealth of interesting algorithmic challenges that can be explored further and constitute a promising area for future work.

Finally, our model provides a basic framework to analyze the phenomena described in the paper in terms of the primitives of the problem represented by the inequality structure, the planner's budget, and the strategic injection of misinformation. Given the salience of these points in modern social learning environments, the model provides a step towards understanding the complex interactions of these factors, and offers guidelines that can help inform policies that aim to reduce inequality and protect society from misinformation.

3.3 Contrasting Bayesian and DeGroot Models

One of the most active areas of inquiry into misinformation examines how the cognitive sophistication of people impacts their ability to fall for misleading content. In this section, using the simplified setting of Mostagir and Siderius (2022a), we capture sophistication by studying how misinformation affects the two canonical models of the social learning literature discussed previously: sophisticated (Bayesian) and naive (DeGroot) learning. We show that sophisticated agents can be *more* likely to fall for misinformation. Our model helps explain several experimental and empirical facts from cognitive science, psychology, and the social sciences. It also shows that the intuitions developed in a vast social learning literature should be approached with caution when making policy decisions in the presence of misinformation. We conclude by discussing the relationship between misinformation and increased partisanship, and provide an example of how our model can inform the actions of policymakers trying to contain the spread of misinformation.

3.3.1 Model

There is a true and unknown state of the world $\theta \in \{L, R\}$ indicating whether a left-leaning or right-leaning idea is correct. There are N agents in the population who are trying to learn θ to make an informed binary decision. Agents receive independent information (messages) about θ and then interact with others to try and learn what the true value of θ is.

Timing. We present our model using a parsimonious three-period model $t = 0, 1, 2$:

(i) At $t = 0$, agents start out with heterogeneous ideological beliefs. The initial belief, $\pi_{i,0}$, of each agent i is drawn i.i.d. from a continuous distribution H, corresponding to her belief in $\theta = R$. Agents with beliefs $\pi_{i,0} < 1/2$ are (initially) left-leaning, while those with beliefs $\pi_{i,0} > 1/2$ are (initially) right-leaning.

(ii) At $t = 1$, agents receive (independent) messages advocating for either state L or state R. Formally, each agent i receives a single message $m_i \in \{L, R\}$.[29] Some of the messages come from *organic news*, which are correlated with the state; in particular, $\mathbb{P}[m_i = \theta] = p > 1/2$ for every agent i. We assume throughout that the population N is large (i.e., $N \to \infty$) so that in the presence of organic news only, the truth is discernible with high probability by aggregating all of the messages.[30]

The messages that agents receive could also be *misinformation* that is orthogonal to the state. Agents cannot discern whether a given message contains misinformation or not. The probability that a message contains misinformation is $q < 1/2$, i.e., most news is organic. Agents are aware of the existence of misinformation (they know q), but do not know how this misinformation is broken down along the two possible states, i.e., they do not know the proportion of misinformation arguing for L vs. the proportion of

[29] Receiving a single message is without loss of generality. Appendix B.4 considers a generalization where agents get multiple (independent) messages over time, but our results apply identically.

[30] In Appendix B.5.2, we measure the sensitivity of our results to this assumption via simulations with finite N populations.

misinformation arguing for R. This follows the empirical observation in van der Linden et al. (2020) that shows that while people agree about the existence of misinformation and even its extent, they do not agree on whether this misinformation leans more left or right.

Let r denote the proportion of misinformation *on the right*. We assume that r is drawn from a differentiable distribution $r \sim F(\cdot)$ at $t = 0$ and is independent of θ. We assume that F has full support[31] on some interval $[\underline{r}, \bar{r}]$ and no support[32] outside of this interval.[33] Similarly, we assume the distribution of prior beliefs H has full support over $[\underline{\pi}, \bar{\pi}]$ and no support outside of it. In other words, we assume the supports of all distributions are *convex*.

(iii) At $t = 2$, agents observe the broadcasted beliefs of other agents in periods $t = 0$ and $t = 1$ and use these beliefs (as well as their own message) to form a final belief $\pi_{i,2}$ (as described below). Following this, each agent i makes a binary decision $a_i \in \{L, R\}$ based on which state she believes is more likely.

We consider two types of populations: **Bayesian** and **DeGroot**. Bayesian agents learn about θ by updating their beliefs in a fully Bayesian way, whereas DeGroots use simple learning heuristics. We use $\mathbf{1}_{\theta=R}$ to denote the indicator function of $\theta = R$ (i.e., $\mathbf{1}_{\theta=R}$ is equal to 1 when $\theta = R$ and 0 when $\theta = L$). Recall $\pi_{i,0}, \pi_{i,1}$, and $\pi_{i,2}$ are the beliefs of agent i at times $t = 0, 1$, and 2, respectively.

(i) **Bayesian Society**: At $t = 1$, each Bayesian agent forms a posterior update about the state, $\pi_{i,1}$, given the article with message m_i and knowing content may contain misinformation:

$$\pi_{i,1}(m_i = R) = \mathbb{E}[\mathbf{1}_{\theta=R} | m_i = R] = \int_0^1 \frac{(p(1-q) + qr)\pi_{i,0}}{p(1-q)\pi_{i,0} + (1-p)(1-q)(1-\pi_{i,0}) + qr} f(r)\, dr$$

$$\pi_{i,1}(m_i = L) = \mathbb{E}[\mathbf{1}_{\theta=R} | m_i = L] = \int_0^1 \frac{((1-p)(1-q) + q(1-r))\pi_{i,0}}{(1-p)(1-q)\pi_{i,0} + p(1-q)(1-\pi_{i,0}) + q(1-r)} f(r)\, dr$$

At time $t = 2$, agents form Bayesian posterior estimates about the state, $\pi_{i,2}$, given their article with message m_i and the beliefs of agents in the population $\{\pi_{j,0}, \pi_{j,1}\}_{j \neq i}$, again, fully aware that there may be misinformation in the system. This is akin to the updating

[31] We define full support of a distribution G on an interval $[a, b]$ as having its density g satisfy the following property: there exists $\mu > 0$ such that $g(\alpha) > \mu$ for all $\alpha \in [a, b]$.

[32] No support over a set $\mathcal{A}$ means the distribution G draws an element from $\mathcal{A}$ with probability 0.

[33] While this does not rule out full support of F on $[0, 1]$, this more general assumption allows us to capture the effect of relatively symmetric vs asymmetric prevalence of misinformation.

process in Acemoglu et al. (2016), where agents are uncertain about the underlying message distribution.

(ii) **DeGroot Society**: DeGroot agents are boundedly rational agents who use a learning heuristic to learn θ. At $t = 1$, each DeGroot agent updates her belief of the state using Bayes' rule taking the news at *face value* (i.e., assuming there is no misinformation in the system). This is similar to how these agents update their beliefs in Jadbabaie et al. (2012)):

$$\pi_{i,1}(m_i = R) = \mathbb{E}[\mathbf{1}_{\theta=R}|m_i = R,\, q = 0] = \frac{p\pi_{i,0}}{p\pi_{i,0} + (1-p)(1-\pi_{i,0})} \tag{3.4}$$

$$\pi_{i,1}(m_i = L) = \mathbb{E}[\mathbf{1}_{\theta=R}|m_i = L,\, q = 0] = \frac{(1-p)\pi_{i,0}}{(1-p)\pi_{i,0} + p(1-\pi_{i,0})} \tag{3.5}$$

Based on these observations, each DeGroot takes an average of all the time $t = 1$ beliefs of the agents in society to form their time $t = 2$ belief, i.e., $\pi_{i,2} = \frac{1}{N}\sum_{j=1}^{N} \pi_{j,1}$. In other words, DeGroot agents employ "rule-of-thumb" learning to update their beliefs instead of forming a Bayesian posterior belief.

Learning. At $t = 2$, agent i chooses a binary terminal action $a_i \in \{L, R\}$ that minimizes her quadratic loss $\mathbb{E}[(a_i - \mathbf{1}_{\theta=R})^2]$ given her belief, $\pi_{i,2}$. We follow the standard definition of learning (e.g., Acemoglu et al. (2011)) and say that society *learns* if all agents take the correct action ($a_i = \theta$); otherwise, society *mislearns*. In Appendix B.5.1, we explore how our results are affected when characterizing the expected proportion of agents who (mis)learn instead of the classical "all-or-nothing" measure of (mis)learning.

Remark — While we adopt the three-period learning model to most transparently demonstrate the main concepts, richer learning dynamics can be supported without compromising any of the key results. In particular, when agents learn from each others' beliefs over a social network (and thus do not observe all beliefs in the population), our findings generalize, provided there is a longer learning horizon and given some mild assumptions on the network structure. The details of the reduction from more general networked learning to the three-period model are supplied in Appendix B.4.

3.3.2 Illustrative Example

We present an example to show that agents who use simple learning heuristics can learn better than fully-rational agents in the presence of misinformation. For concreteness, we fix $\theta = L$ as the true state (which, by assumption, is unknown to the agents). We also assume the misinformation ideological split r is uniformly distributed on $[0, 1]$, i.e., the split is ex-ante symmetric for left-leaning and right-leaning misinformation. Agents do not know the exact value of r, but they know that it comes from the uniform distribution on $(0, 1)$. We compare the following two settings:

Setting A: Weak Organic Messages and No Misinformation. Consider the baseline case studied throughout the social learning literature. There is no misinformation, i.e., $q = 0$, and $p = 0.54$, so that 54% of organic messages align with $\theta = L$. In this setup, organic news is (weakly) correlated with the truth.

Do agents learn the correct state (almost surely) when the population is large? The short answer, as already developed in a vast literature, is yes. Both agent types correctly learn that the true state is L. This happens regardless of their initial prior beliefs (i.e., even those on the extreme right still learn that the correct state is L) and despite the fact that news is weakly correlated with θ. In this setup, both the Bayesian and DeGroot societies take the correct action.

Setting B: Misinformation with Stronger Organic Messages. In this setting, misinformation exists in the system, with $q = 0.25$. On the other hand, organic news is of higher quality, with $p = 0.6$. For this example, assume that $r = 0.64$, i.e., 64% of the misinformation advocates for $\theta = R$ and 36% advocates for $\theta = L$. This means that among the large collection of messages $\{m_i\}_{j=1}^{N}$, roughly 54% correspond to L (i.e., $p(1-q) + (1-r)q$) and 46% correspond to R (i.e., $(1-p)(1-q) + qr$). This distribution is *identical* to the one in Setting A, under which agents were able to correctly learn. In Setting B, if agents did not know misinformation exists, learning will proceed exactly as before and everyone will take the correct action despite the presence of misinformation. However, agents are now aware that there is misinformation in the system. Given this, we analyze how each society updates its beliefs.

Bayesian agents: Agents observe the initial beliefs $\pi_{j,0}$. At $t = 1$, each Bayesian agent i forms a posterior belief $\pi_{i,1}$ based on m_i as mentioned in Section 3.3.1. Observe that $\pi_{j,1} > \pi_{j,0}$ if and only if $m_i = R$ and $\pi_{j,1} < \pi_{j,0}$ if and only if $m_i = L$. Thus, by observing beliefs in the second period, every agent i can deduce the messages $\{m_i\}_{j=1}^N$. As noted, When N is large, the law of large numbers guarantees that roughly 54% of the messages that agent i observes will favor L.

Note that there are *exactly* two realizations of r for which 54% of the messages are L and 46% are R. The first is the true realization, where $\theta = L$ and $r = 0.64$. The other is where $\theta = R$ but $r = 0.04$ (i.e., only 4% of the misinformation is on the right). This situation is depicted in Figure 3-15. Because r is uniformly distributed, *both* of these scenarios are equally likely. This implies that $\pi_{j,2} = \pi_{j,0}$ because the messages provide no information about the state θ. In other words, right-leaning agents spin a narrative that the vast majority of misinformation is on the left, and cannot use the massive quantity of news to change their views. Similarly, left-leaning agents do the same, believing (correctly, in their case) that most of the misinformation must be on the right. The society of Bayesians does not learn and there is persistent disagreement about θ in the population.

DeGroot agents: DeGroot agents take messages at face value and use them to update their beliefs (via Equations (3.4) and (3.5)) using $p = 0.6$. Thus, noting that 54% of messages are L and 46% of messages are R, DeGroot agents hold beliefs $\pi_{j,t}$ for all $t \geq 2$:

$$\int_0^1 \left(.46 \cdot \frac{.6\alpha}{.6\alpha + .4(1-\alpha)} + .54 \cdot \frac{.4\alpha}{.4\alpha + .6(1-\alpha)} \right) \, d\alpha = 0.495 < 1/2$$

so all agents in society learn the correct state, in contrast to the Bayesians.

Likelihood of Mislearning. We used a specific misinformation split $r = 0.64$ in the above example for simplicity. Other values of r would give rise to different outcomes. Instead of focusing on a specific realization of r, we can instead look at the *likelihood* that society does not learn when the split r is randomly drawn from its true (uniform, in this example) distribution. How do Bayesian agents perform relative to DeGroot agents *on average*?

It turns out that in this case, Bayesian agents mislearn *twice as much* as DeGroots (probability that Bayesians mislearn is 40% vs. 20% for the DeGroots – see Appendix B.1 for calculations). We characterize this ratio as a function of the distribution of r in Theorem 3.3.2. Generally, while Bayesian agents thrive in environments where information is organic, they are much more vulnerable to mislearning and taking the wrong action in the presence of misinformation.

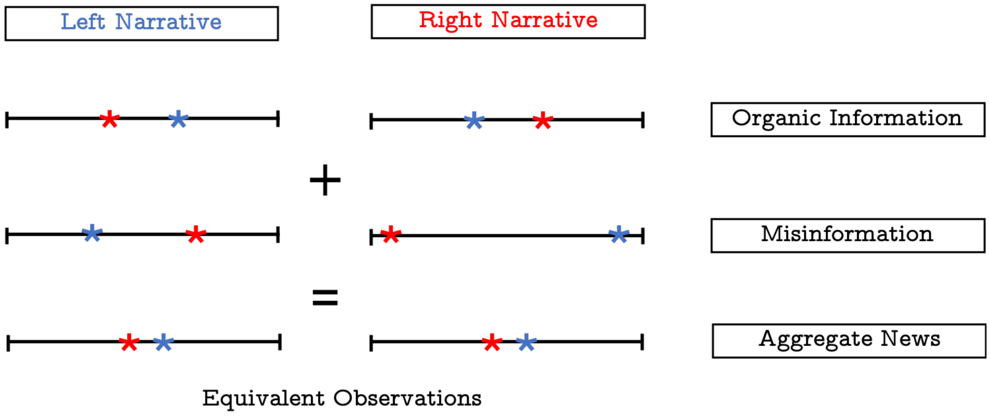


Figure 3-15. Two narratives that give rise to equivalent observations. In this example, the true state is L and most of the misinformation comes from the right with $r = 0.64$, i.e., the left narrative is the correct one. However, agents who hold right-leaning beliefs can rationalize their observations as coming from the right narrative, with the organic information arguing for R and most of the misinformation coming from the left, with $r = 0.04$.

3.3.3 Learning in Bayesian vs. DeGroot Societies

In this section, we generalize the previous example and investigate the conditions under which learning breaks down in each society. We follow this up with two technical results. Theorem 3.3.1 characterizes *when* Bayesians learn worse than DeGroots as a function of the *amount* of misinformation in the system, and Theorem 3.3.2 quantifies *how much* worse they learn as a function of the *distribution* of that misinformation.

DeGroot (Mis)learning: Propaganda for the Incorrect State

DeGroot agents update their beliefs about the state by averaging the opinions of others. Because of this simple updating process, they always converge to a (possibly incorrect) consensus about what the true state is. When there is no misinformation (i.e., $q = 0$), agents *will* learn the correct state and choose $a_i = \theta$ (see Golub and Jackson (2010)). When misinformation is present, we can provide necessary and sufficient conditions for DeGroot learning in terms of the ideological split of misinformation r:

Proposition 3.3.1. *When $\theta = L$, there exists a threshold r_D^* such that if $r < r_D^*$, the DeGroot society learns and if $r > r_D^*$, the DeGroot society mislearns.*

Proposition 3.3.1 shows that failure of learning depends on whether propaganda for the incorrect state is sufficiently high to direct agents away from the belief that the organic news argues for. Agents learn as long as this propaganda does not overpower organic news.

Bayesian (Mis)learning: Rationalization and Competing Narratives

Bayesian agents make inferences about the state by observing the distribution of messages in the population. They then rationalize the values for the misinformation split r that gives rise to this message distribution. We refer to these values as *narratives*. There can be at most two narratives: one that corresponds to $\theta = L$ and one that corresponds to $\theta = R$. Only one of these narratives is correct, and agents will stick with the narrative that fits their beliefs.[34] However, there can also be a *single* correct narrative, in which case all agents learn the true state θ. This again depends on the extent of misinformation arguing for the incorrect state:

Proposition 3.3.2. *When $\theta = L$, there exists a threshold r_B^* such that if $r < r_B^*$, a single narrative exists and the Bayesian society learns.*

If $r > r_B^*$, then two narratives exist. Figure 3-16 shows this situation for a specific $r > r_B^*$ value. Recall that agents do not know r, but in this example they know that it follows a triangular distribution. This makes them believe that the more likely narrative is the one corresponding to R. In this case, *all agents* move further to the right (and away from the true state). Note that unlike the example in Section 3.3.2, where the two competing narratives were equally likely under the uniform distribution, the triangle distribution makes the incorrect narrative strictly more likely. Section 3.3.3 shows how to quantify the likelihood of mislearning as a function of the hazard rate of the distribution of r.

Low vs High Misinformation Regimes

Next, we present our main theorem, which analyzes the settings under which DeGroot or Bayesian agents mislearn more often:

Theorem 3.3.1. *Suppose H is symmetric about $1/2$.*[35] *Then there exists a threshold $q^* \in (0, 1)$ such that:*

[34] This is reminiscent of the effect informally described by the Today show co-host Al Roker, commenting on conflicting results of science experiments: “I think the way to live your life is you find the study that sounds best to you and you go with that."

[35] This corresponds to a society where every belief on the left is perfectly mirrored by a belief on the right of the same extremity. This allows us to analyze both the $\theta = L$ and $\theta = R$ cases identically and abstracts away from scenarios where society begins either initially biased toward or away from the correct state, in order to focus on the underlying learning mechanisms.

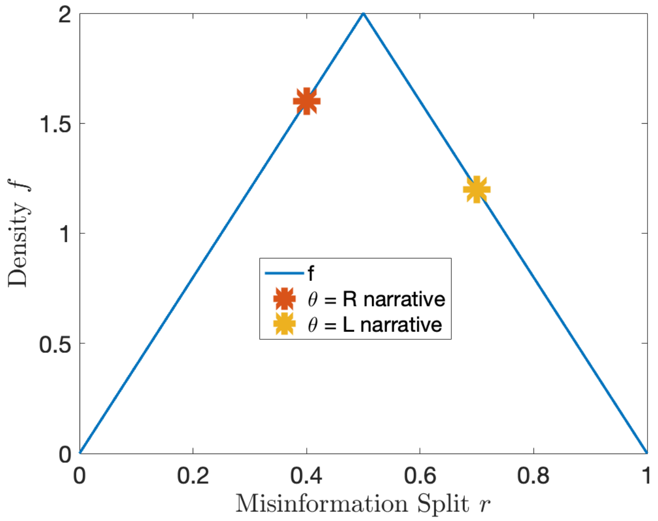


Figure 3-16. Two narratives that justify the same observed message distribution under a specific misinformation split r. Agents do not know r but they know that it follows a triangular distribution. This makes them believe that the more likely narrative is the one that corresponds to the true state being R. In this case, all agents move further to the right and away from the correct state.

(a) If $q < q^*$, the Bayesian society mislearns with lower probability than the DeGroot society;

(b) If $q > q^*$ and H and F have full support on $[0, 1]$, the Bayesian society mislearns with strictly higher probability than the DeGroot society.

Theorem 3.3.1 establishes that when the amount of misinformation is not too large, Bayesian agents can still learn better than their DeGroot counterparts. This is the classic intuition from the social learning literature and the one empirically observed in Bronstein et al. (2019); Bago et al. (2020); Pennycook and Rand (2019). On the other hand, once misinformation becomes more rampant, DeGroot agents become *more* adept at aggregating the organic information compared to the Bayesians, who find themselves in consistent disagreement over what the truth is, as in the empirical work of Drummond and Fischhoff (2017); Kahan et al. (2012); Hamilton et al. (2015).

The intuition for Theorem 3.3.1 is as follows. When misinformation is relatively low, the organic news is enough for Bayesians to infer the true narrative and see the misinformation as purposefully deceptive. On the other hand, with DeGroot agents, this misinformation still has a chance of successfully steering the beliefs of the population away from the truth. Once the amount of misinformation becomes large, a *post-truth* effect kicks in for the Bayesians: any reasonable narrative can be told about the source of misinformation, and disagreement over the true state ensues. While DeGroots are not guaranteed to learn either, their simple updating tends to work well when misinformation is not too heavily skewed in one direction or another. Thus, in more instances, the DeGroot society is able to learn from the organic news while allowing the misinformation on both sides to nearly "wash out." We can quantify how much better the DeGroots do in this case, which is the subject of the next section.

High Misinformation and Mislearning Rates

When $q > q^*$, we observe in Theorem 3.3.1 that DeGroot agents learn more effectively than Bayesian agents. This section is for readers who are interested in quantifying the extent with which DeGroots outperform Bayesians. This depends on the misinformation split distribution $F(\cdot)$. We formalize this as follows. The exact difference in rates of mislearning between Bayesian and DeGroots in the high-misinformation regime (i.e., $q > q^*$) is determined by the *hazard rate* $\lambda_F(\alpha)$ of F. Recall that the hazard rate is given by $\lambda_F(\alpha) = \frac{f(\alpha)}{1-F(\alpha)}$, where $f(\cdot)$ is the density of cumulative distribution function F. Denote by μ the the relative frequency of DeGroot to Bayesian mislearning (which by Theorem 3.3.1 is less than 1), then we can characterize how μ changes as a function of the strength of organic signals p (i.e., whether agents can learn from strongly informative content):

Theorem 3.3.2. *Suppose that H and F have full support on $[0, 1]$, H is symmetric, and $q > q^*$ as in Theorem 3.3.1(b). Consider $\alpha = \frac{1-2(1-p)(1-q)}{2q}$ and $\beta = p\left(1 - \frac{q^*}{q}\right)$.*[36] *The ratio μ is increasing in p if $\lambda_F(\alpha) < 2\lambda_F(\alpha - \beta)$, decreasing in p if $\lambda_F(\alpha) > 2\lambda_F(\alpha - \beta)$, and unchanging if $\lambda_F(\alpha) = 2\lambda_F(\alpha - \beta)$.*

Informally, Theorem 3.3.2 states that Bayesians perform comparatively worse than DeGroots when misinformation is more evenly distributed. The reason is that while Bayesians are adept at making inferences about the possibility of strongly misleading misinformation, misinformation that is relatively balanced on both sides permits more rationalization of narratives and more disagreement. On the other hand, more balanced misinformation is always better for DeGroot learning because it permits greater likelihood of *overall balanced* news and less propaganda for the incorrect state. Appendix B.3 shows how to apply this result when r comes from skewed or unskewed distributions.

[36] Note that it can be shown $0 \leq \alpha - \beta \leq \alpha \leq 1$, so the hazard rates are well-defined everywhere.

3.3.4 Polarization

Polarization and political partisanship have been steadily increasing (see Abramowitz (2010) and Pew Research Center (2014) for evidence from the United States). As noted in the introduction, a substantial literature advocates that increased polarization makes people more likely to disagree over objective facts because of politically-biased reasoning, whereas new findings (see Tappin et al. (2020)) suggest that this disagreement can be explained as a byproduct of sophisticated (Bayesian) reasoning and not partisan bias.

Our model lends support to the latter explanation by showing that disagreement can naturally arise from Bayesian updating. More generally, Theorem 3.3.3 highlights the fact that attempts to establish a connection between partisanship and disagreement should take into account the sophistication level of the agents and the amount of misinformation in the system. As misinformation becomes more prevalent, increased polarization leads to failure of learning in Bayesian societies, but leaves DeGroot societies relatively unaffected. Thus, measuring the effects of polarization on learning and consistent disagreement without incorporating these elements can lead to seemingly contradictory evidence.

We operationalize polarization as follows: consider some symmetric (about $1/2$) belief distribution H (density h) with support $[\underline{\pi}, \bar{\pi}]$ and a mean-preserving spread of H to some $\tilde{H}_\gamma$ (density h_γ) defined as:

$$\tilde{h}_\gamma(\pi + (\pi - 1/2)\gamma) = \frac{1}{1+\gamma} h(\pi)$$

where $\gamma \in [-1, \bar{\gamma}]$ where $\bar{\gamma} = \min\left\{\frac{1-\bar{\pi}}{\bar{\pi}-1/2}, \frac{\pi}{1/2-\underline{\pi}}\right\}$. One can think of γ as a measure of belief polarization in society. For larger γ, $\pi + (\pi - 1/2)\gamma$ is closer to 0 when $\pi < 1/2$ and closer to 1 when $\pi > 1/2$; thus, the probability of realizing "tail" beliefs grows when γ increases. (Note that $1/(1+\gamma)$ is simply a scaling factor to guarantee the $\int_0^1 \tilde{h}_\gamma(\pi)\, d\pi = 1$.) A simple example of increasing polarization for a truncated normal distribution of beliefs is given in Figure 3-17.

Our next result establishes a threshold result for polarization and its consequences to both Bayesian and DeGroot mislearning:

Theorem 3.3.3. *Let H be symmetric about $1/2$ and $q > q^*$. There exists a threshold γ^* such that if $\gamma < \gamma^*$, the DeGroot society mislearns more often than the Bayesian society, whereas if $\gamma > \gamma^*$, the DeGroot society mislearns less often than the Bayesian society.*

In Bayesian societies, an increase in belief polarization always hurts learning. Mean-

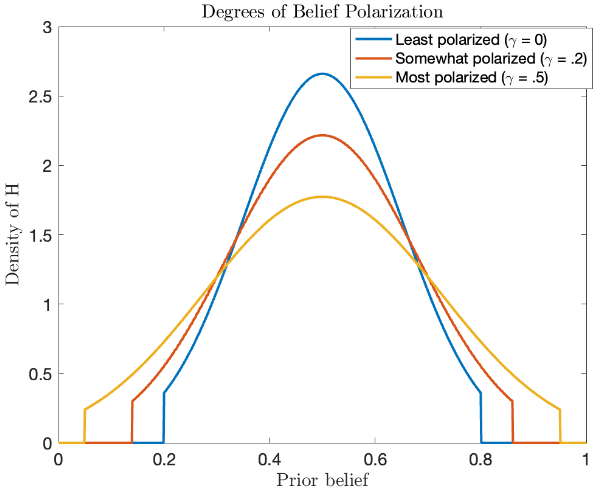


Figure 3-17. (Color online) Polarization of beliefs captured through parameter γ.

preserving spreads necessarily create more tension in the learning process because agents cannot agree on what is likely to be the true narrative. The impact is more mild on DeGroot agents because they communicate and update beliefs by taking averages of their neighbors. In this case, mean-preserving spreads do not affect their general ability to aggregate organic information, even when they start from very different initial beliefs.

Remark — Note that when H is asymmetric, polarization still always hurts Bayesian societies (see the proof of Theorem 3.3.3 in Appendix A). However, it is possible that in this case polarization *improves* the chances of learning in a DeGroot society (see Appendix B.2 for an example). This occurs because when society is initially well-informed, additional polarization pulls initial opinions towards the correct and incorrect states. However, the convexity of how agents update their beliefs from news (see Equations (3.4) and (3.5)) leads to an overall movement towards the correct state. This fails in Bayesian societies because of persistent disagreement about the truth, and further accentuates the result of Theorem 3.3.3 that polarization is more damaging to a Bayesian society than a DeGroot one.

Targeting Policies. We now consider the problem of a planner who can target a subset of the population with information arguing for one state over the other. One can think of this policy as an educational outreach intervention. For example, governments have been ramping up their efforts to convince citizens to vaccinate against COVID-19, and a part of these efforts is targeted advertising. The question is which agents should the planner target? For example, should she target the most polarized agents? We show that the answer, again, crucially depends on the level of sophistication of the agents.

Formally, the planner informs a small but positive measure of agents that the correct state is either L or R. We assume the agents interpret this information in the same way they do news: combined with the message m_j that agent j receives, she also gets the planner's message $m_j^p \in \{L, R\}$ and updates using both messages. The next result describes the planner's targeting policy:

Proposition 3.3.3. *There exists $1/2 < \pi^* < 1$ such that the targeting policy for DeGroot agents is to target those agents whose beliefs lie in an open interval containing π^* (and bounded away from 1). The policy for Bayesians agents is to target those agents whose beliefs are farthest from the truth, i.e., the extremists.*

Proposition 3.3.3 states that in a DeGroot world with $\theta = L$, the planner wants to influence right-leaning *moderates*, as these are the agents who change their belief most when seeing message L. Extremists in this society mostly dismiss messages that don't agree with their priors, and the planner has little to gain by targeting them. However, with Bayesian agents, *extremists* are exactly the agents that the planner needs to target. While the efficacy of her work is limited, these are the agents who are most inclined to spin narratives that anchor them to the incorrect state.

Remark 1 — Proposition 3.3.3 recommends a policy when the planner knows the true state, but the insights generalize when the planner herself is uncertain about what the state actually is. In a DeGroot society, the planner tries to make relatively moderate agents (of both ideologies) more moderate. In Bayesian societies, the planner tries to make the extremists (of both ideologies) move toward the center. Both of these policies push toward decreasing the polarization of ideological beliefs (albeit in different ways).

Remark 2 — Under a different learning objective, such as minimizing the proportion of agents who mislearn the state, the DeGroot policy in Proposition 3.3.3 remains unaffected, but the Bayesian policy becomes more subtle. In many environments, the regulator should still target the most polarized Bayesian agents in society, but in some other instances, targeting mostly moderate Bayesian dissenters can be more effective. The nuances of targeting interventions under this alternative objective are explored in detail in Appendix B.5.1.

3.3.5 Final Remarks

As argued recently in Watts et al. (2021), accurate information is very much a prerequisite for successful democratic discourse. The Internet and social media have made it easier to disseminate misinformation (Allcott and Gentzkow, 2017), with far-reaching consequences.[37] There are ongoing efforts across multiple disciplines to try and uncover the mechanisms by which misinformation spreads. In this paper, we contribute to these efforts by examining misinformation through the lens of social learning and focusing on agents' sophistication types. While the learning mechanisms of these types have been studied in a broad context, they have not been analyzed and compared when there is rampant misinformation. We show a reversal of results and intuitions that hold in many normative learning setups, but not in the presence of misinformation. We do this through a parsimonious framework whose results reconcile several empirical studies and whose predictions show the need for researchers and policymakers to jointly consider sophistication and social learning as integral components in studying the spread of misinformation.

[37] Examples range from individual actions along the lines of Pizzagate (Fisher et al., 2016), to belief in the "Death Panels" of the Affordable Care Act (Watts et al., 2021), to more collective action failures like the spread of measles in Eastern Europe as a result of Russian disinformation (Broniatowski et al., 2018).

Chapter 4

Attention Models and Algorithmic Ranking

In this chapter, we address topics related to user attention, content creation (and engagement), and platform algorithms. In Section 4.1, we provide and analyze a model of content creation with limited user attention to characterize how users allocate their time to various articles when there is a deluge of online content. Using this framework, we study the incentives for content creation in the digital era, when technology makes it increasingly easier to supply content. In Section 4.2, we experimentally investigate the behavioral decisions users make to click on and share certain articles. We use this to engineer a design to test some natural platform ranking algorithms, such as preference-based or friend-based, and gauge user behavior in response to such changes. In particular, we are interested in how these algorithms might promote engagement with certain types of content such as misinformation or catchy click-bait.

4.1 Competing for Limited Attention on Social Media

We consider a model of endogenous content creation on a social media platform. Articles consist of two attributes, catchiness and informativeness, one of which is an easily observed characteristic and the other which is hidden and may only be learned through sufficient reading. A representative platform user with limited attention receives utility from both entertainment and from absorbing information, and endogenously chooses how to allocate her time reading based on the headlines she observes. We fully characterize the equilibria when there is a single article, and show there can be multiplicity because of strategic complementarity between the user (who wants to read informative content) and the provider (who only wants to invest if the content will be read). We generalize this characterization to a monopolist who has access to a technology that can produce many articles, which admits a much richer set of equilibria. As the number of articles grows, we find that while the most informative equilibrium supplies more information, the least informative equilibrium also supplies less information. Once competition is introduced across content providers (holding the total number of articles fixed), information provision deteriorates in both the most information rich and least information rich equilibria, because providers cannibalize each other through catchy content to steal user attention, but which provides little informational value.

4.1.1 Model

We consider a single consumer who processes content on a social media site with N content providers. There is a multi-dimensional state $\boldsymbol{\theta}^* \equiv \{\theta_1^*, \ldots, \theta_K^*\} \in \mathbb{R}^K$, and each component θ_k^* of the state vector $\boldsymbol{\theta}^*$ is drawn i.i.d. from a standard normal distribution. The representative consumer digests content on social media potentially for both entertainment and to get a better understanding of the world (a tighter estimate of $\boldsymbol{\theta}^*$).

Content provides utility through both direct and indirect channels. Content which is *catchy* is more interesting to read, and provides some fixed, direct utility regardless of any other factors. On the other hand, content that is *informative* provides indirect utility by allowing the consumer to form a better estimate of $\boldsymbol{\theta}^*$. Each article has a two-dimensional type describing whether it is (i) catchy or non-catchy and (ii) informative or non-informative. While users can observe catchiness on the surface, they cannot gauge informativeness directly without reading the article.

Attention Game. There are three stages to the game. In stage 1, content providers create content for users to consume. In stage 2, content consumers inspect (or "survey") the set of articles available to them based solely on their headlines (and catchiness). In stage 3, consumers dynamically allocate time to different articles. These stages are detailed next:

(1) *Content creation*: There are N content providers who invest in creating (at most) M pieces of content each in stage 1 (which we call "articles"). First, each provider selects up to M topics (not necessarily distinct) to write articles on. Second, each provider decides whether to invest in the catchiness and / or informativeness of an article. Both such investments

are costly.

For each article, there is a fixed investment $C > 0$ for each of provider j's articles to make it catchy ($\mathcal{C}$) instead of non-catchy ($\mathcal{C}^{\circ}$). For each article a written by producer j, there is a private investment cost $I_{j,a} > 0$ to produce informative content ($\mathcal{I}$); non-informative content ($\mathcal{I}^{\circ}$) is costless. We assume $I_{j,a}$ is drawn from a smooth distribution F with full support over $[0, \bar{I}]$ for some $\bar{I} > 0$. These costs are independent across providers j and articles a, and privately known by the provider.

If a provider elects to supply informative content on topic k, it will eventually reveal information about state θ_k if the consumer reads for long enough. Formally, an informative article a from provider j (written on topic k) will generate a signal $s_a \sim \theta_k + \varepsilon_{j,a}$, which arrives at some random time after the consumer has been reading (as described below). It is assumed $\varepsilon_{j,a}$ is unbiased Gaussian noise with variance σ^2 i.i.d. across content providers and articles. Uninformative content never provides any signal. Content creators receive a payoff proportional to the time the consumer spends reading their content, less any investments into catchy or informative content. In other words, provider j receives payoff V_j given by:

$$V_j = \sum_{a \in \mathcal{A}_j} t_a - C\mathbf{1}_{\mathcal{C}(a)=\mathcal{C}} - I_{j,a}\mathbf{1}_{\mathcal{I}(a)=\mathcal{I}}$$

where $\mathcal{A}_j$ is the set of articles produced by provider j, t_a is the time the consumer spends reading the article, $\mathcal{C}(a)$ is the catchiness type of article a, and $\mathcal{I}(a)$ is the informativeness of article a.

(2) *Consumer inspection*: In stage 2, while the consumer cannot read all articles in detail, she can garner some information about each based on the article's "headline". Formally, we assume each article has an easily digestible headline indicating (i) the topic k of the article and (ii) whether the article is catchy ($\mathcal{C}$) or non-catchy ($\mathcal{C}^{\circ}$). That is, for every article a, the catchiness type $\mathcal{C}(a)$ is revealed to the consumer. However, the headline provides no direct information about whether the article is informative; in other words, $\mathcal{I}(a)$ is unobservable for each article a.

(3) *Consumer reading*: In stage 3, we consider a continuous time environment $t \in [0, \infty)$ where the consumer dynamically chooses which articles to read and for how long.

Catchy articles offer a bulk utility of κ from entertainment conditional on the consumer reading for at least time $\tau_C > 0$. Non-catchy content provides no (direct) utility. The consumer has an outside option that provides constant utility flow $v\,dt > 0$ and can be collected for every time interval dt if she chooses to abstain from reading any content (where we assume $\kappa > v\tau_C$).[1]

Consumers cannot tell which articles are informative solely from the headline. However, informative articles always generate a *reputability signal* (after the consumer reads for awhile) followed by an *information signal* that gives the consumer valuable information about state θ_k. Uninformative articles never provide any signals. Precisely, an informative article generates a noiseless binary reputability signal that the article is quality (or not) after some random time determined by a Poisson clock with parameter $\lambda_I > 0$. This is always followed by an *information signal* s_a after additional reading time $\tau_I > 0$ about state θ_k.[2]

Because the consumer has limited attention, the reading process eventually terminates. This is determined by a Poisson clock with parameter λ; we denote T as the random time at which the attention game ends and payoffs are realized. At time T, the consumer takes an action $\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} \mathbb{E}[||\boldsymbol{\theta} - \boldsymbol{\theta}^*||_2^2]$, which provides utility $-\beta||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*||_2^2$ for some parameter $\beta > 0$.

Formally, at time T, the representative consumer receives utility:

$$U = vt_o - \beta||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*||_2^2 + \sum_{j=1}^{N} \sum_{a \in \mathcal{A}_j} \kappa \mathbf{1}_{t_a > \tau_C}$$

where t_o is the time spent on the outside option and t_a is the time spent on article a.

Equilibrium Concept. We consider sequential equilibria of the attention game between N content providers and a representative consumer. In other words, each supplier trembles with vanishingly small probability $\varepsilon > 0$ for both catchiness and informativeness. This refines

[1] If $\kappa < v\tau_C$, the reader finds the outside option more appealing than a catchy article, leading to a less rich equilibrium structure, but one which still follows easily from our analysis.

[2] This can be interpreted as a reader who, after investing some initial time reading the article, can discern whether the article is reputable, but must then read to the end to get the full informational value. The uncertainty about the arrival of the reputability signal models the consumer's uncertainty in how much of the article she must first read before knowing whether continuation is worthwhile or not.

equilibria in which all provider strategies are to supply one type of headline (either catchy or non-catchy), and deviations from this leave consumer beliefs entirely unstructured.

Optimal Reading Scheme. We make a few parameter assumptions that are used throughout the remainder of the paper. First, we assume that $\beta(\lambda + \lambda_I)/(1 + \sigma^2) > v(1 + \tau_I)$. This implies that a consumer who knows an article contains information should always continue reading instead of abandoning in favor of her outside option. Second, we assume that $\tau_C > C$. This guarantees that a reader will spend enough time on a catchy article (assuming this is her only option) to make the investment in catchiness profitable for the provider. Under these assumptions we obtain the general structure of a consumer's optimal reading scheme:

Lemma 4.1.1. *In every equilibrium, the consumer's reading strategy always takes the following form. There is an ordered (NM)-tuple of articles $(i_1, \dots, i_{NM})$ and stopping times $(\tau^{(1)}, \dots, \tau^{(NM)})$ such that the consumer reads articles sequentially, stopping after time $\tau^{(i)}$ for article i unless a reputability signal has arrived (in which case she reads for τ_I longer after this signal).*

Lemma 4.1.1 drives provider incentives for supplying different types of content. Consumers (with limited attention) quickly jump between articles, allocating only a short amount of time $\tau^{(i)}$ to each article i before moving onto the next, unless hooked by the information contained. However, it is not possible for a provider to credibly reveal whether an article contains valuable information through just its headline. This will be the driving force behind the kind of content provided in equilibrium.

4.1.2 Monopolist Content Provider with a Single Article

In this section, we consider a monopolist content provider ($N = 1$) with a single article ($M = 1$). The consumer can either interact with the article supplied by the monopolist or allocate her attention elsewhere (i.e., to her outside option). A monopolist producing a single article makes a single choice among four alternatives: $\{\mathcal{C}, \mathcal{I}\}$, $\{\mathcal{C}, \mathcal{I}^\circ\}$, $\{\mathcal{C}^\circ, \mathcal{I}\}$, $\{\mathcal{C}^\circ, \mathcal{I}^\circ\}$ (i.e., deciding on both catchiness and informativeness). If the provider invests in catchiness, then he is guaranteed a reading time of at least $\min\{\tau_C, T\}$. If the provider invests in an informative article, but which is non-catchy, then the consumer would only read to acquire an information signal. However, the consumer does not know if the article is informative or not. If the consumer were to receive a reputability signal (indicating the article is informative), she will continue reading until she acquires the full information necessary to update about the state. Thus, the provider's investment in informativeness depends on the likelihood that the consumer will read for some time before making her assessment about the article's quality, which is endogenously determined in equilibrium.

Equilibrium Characterization

Our next result characterizes the set of equilibria for a monopolist with a single article.

Theorem 4.1.1. *There exist* $0 < \underline{\tau} < \bar{\tau} < \infty$ *such that:*

(a) If $\tau_C > \bar{\tau}$*, there is a unique equilibrium where the provider supplies a catchy article;*

(b) If $\underline{\tau} < \tau_C < \bar{\tau}$*, there exists a unique long-read equilibrium with non-catchy but possibly informative content. In this equilibrium, the consumer has a high stopping time (*τ_H*) and the provider has a high likelihood of providing informative content (*p_H*);*

(c) If $\tau_C < \underline{\tau}$*, there exist two equilibria, the long-read equilibrium and the short-read equilibrium with non-catchy but possibly informative content. In this equilibrium, the consumer has a low stopping time (*τ_L*) and the provider has a low likelihood of providing informative content (*p_L*);*

where $\tau_L < \tau_H$ *and* $p_L < p_H$.

There are three regimes that emerge from Theorem 4.1.1. In regime (a), catchy content will engage a reader for a long time; in this case, the provider is best off providing catchy click-bait, which the consumer reads with little anticipation of information. In regime (b), catchy content provides transient entertainment and consumers tend to move on quickly. In this case, information is supplied with some high probability p_H but the content is non-catchy; the consumer spends a substantial amount of time reading τ_H with the hope that the article will deliver information. In regime (c), when catchy content does not engage the user for long at all, multiple equilibria arise. In particular, a second equilibrium emerges in addition to the equilibrium of (b) where information is supplied with a low probability p_L and the consumer reads for a short period τ_L. This multiplicity occurs due to strategic complementarity between the consumer and the provider: consumers prefer to read content likely to be quality for longer, and suppliers are more willing to invest in such content if they expect to engage the reader for longer.

The intuition for Theorem 4.1.1 can be seen in Figure 4-1. We first define $\Delta(\tau)$, which is a function of an endogenously determined optimal stopping time τ for the reader. This corresponds to the *added* reading time a quality investment expects to receive given the reader employs a stopping time τ. Note that Δ is first increasing in τ, then is decreasing. Short attention for non-catchy articles makes it unlikely for the reputability signal to arrive, and offers little chance that the reader will engage more conditional on a quality investment. On the other hand, for very long attention spans, the consumer has a higher likelihood of quickly getting the information from a quality article, reading to completion, and moving on. An article with no information that strings the reader along for awhile can even be *more* profitable than one with information for extremely long attention spans. This effect drives the eventual decrease in Δ.

Second, we define $\mathcal{L}(I^*)$, corresponding to the best-response stopping time a reader would employ if the content provider invests in quality whenever $I < I^*$ is realized. Observe that $\mathcal{L}$ is monotonically increasing in I, because as the prior probability of quality increases, the consumer is willing to read for longer even in the absence of a reputability signal. The candidates for mixed equilibria are exactly the stopping times τ where $\Delta(\tau) = \mathcal{L}^{-1}(\tau)$; that is, where the additional reading time from the consumer conditional on quality matches the investment cost to the provider.

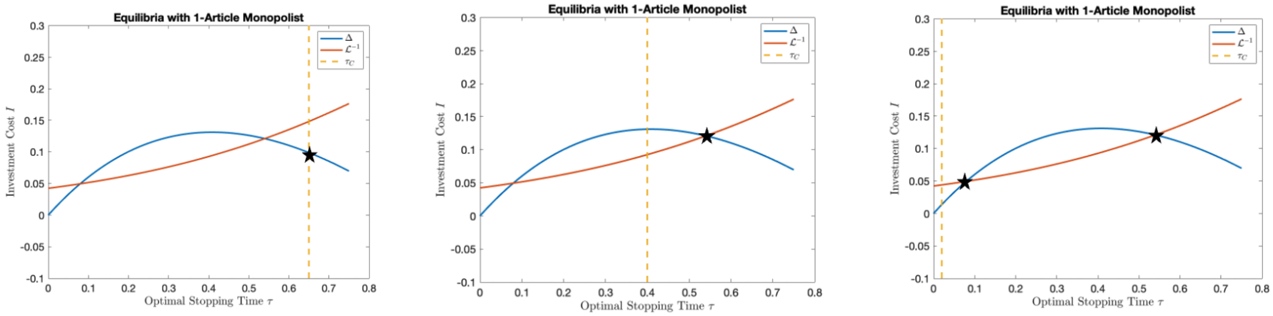


Figure 4-1. Regime (a), regime (b), and regime (c), respectively. Stars indicate equilibrium points and the investment cost threshold I^* such that those providers with $I < I^*$ invest in informativeness and those with $I > I^*$ do not.

In regime (a), catchy articles capture the consumer's attention so much that the provider can gain more from investing in catchiness than informativeness. In this case, there may be still some investments in quality (with probability $\max\{0, \Delta(\tau_C)\}$) but the reader's stopping time is chosen solely absorb the entertainment. In regime (b), there is no catchiness and a unique equilibrium with a high stopping time and high likelihood of information; the optimal stopping time $\tau^{(1)}$ is determined by the higher intersection of Δ and $\mathcal{L}^{-1}$. The lower intersection is not an equilibrium in regime (b) because the provider would find it more profitable to get additional reading time via catchiness than accept the low stopping time. In regime (c), there are two equilibria, determined by both the low and high intersections of Δ and $\mathcal{L}^{-1}$.

Comparative Statics

We define the *most information rich* equilibrium to be the one where the expected improvement in learning of $\boldsymbol{\theta}$ in equilibrium, given by $\mathbb{E}_{\boldsymbol{\theta}}[\mathbb{E}[||\boldsymbol{\theta}||_2^2] - \mathbb{E}[||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_2^2]]$, is maximal. A sufficient condition for an equilibrium to be the most information rich is that the probability of investment in quality is largest and the reader's stopping time is largest for each article (of any equilibrium).[3] In the next result, we provide some straightforward supply-side comparative statics.

Proposition 4.1.1. *The environment is more information rich whenever:*

(i) Information is more easily produced (i.e., $F' \preceq_{FOSD} F$);

(ii) Catchy content is more difficult to produce (i.e., $C' > C$).

The comparative statics of Proposition 4.1.1 are expected. On the supply side, whether the environment admits a more information rich equilibrium depends on whether supplying information is relatively less expensive than supplying catchy content. In many ways, the techological innovations of digital media have vastly improved the extent to which information can be disseminated, providing a possible pathway for more information rich equilibria per Proposition 4.1.1.

We note, however, that information richness generally admits non-monotone comparative statics on the demand side. This includes comparative statics with respect to general consumer attention (λ) and the time needed to acquire information (λ_I and τ_I). This is due to the non-monotonicity of Δ and a *free-rider* problem with content providers. If the reader is more willing

[3] This condition is especially useful in the one-article case, because it shows that the high-readership equilibrium is always the most information rich equilibrium, unless in regime (a) of Theorem 4.1.1, where the catchy equilibrium is unique. With multiple articles, as we discuss in Section 4.1.3, there may be no equilibrium that has both the highest investments in quality and highest reading times across all articles.

to invest time in an article before seeing a reputability signal, that also means the provider can guarantee a long reading time *even* if he does not invest in quality. To see this formally, consider a long-read equilibrium that intersects Δ on the downslope (as is the case in Figure 4-1). Here, there are diminishing returns from investing in quality as the reader increases her stopping time. When, general attention increases (for example), $\mathcal{L}^{-1}$ shifts to the right, and the reader is willing to spend more time experimenting with a non-catchy article to see if it is quality and contains information. She thus employs a longer reading time before needing to see any signal of reputability. However, this creates incentives for providers to supply overall worse quality content, for instance, via longer meandering articles that are on average less likely to convey anything of value.

Sometimes, the opposite comparative statics can hold. This can happen, for instance, when general attention was previously quite short but is now longer. The comparative static will be reversed when the long-read equilibrium occurs at an intersection point of Δ on the upslope. At this intersection, content providers a more likely to invest in quality with a longer reading time, because it gives them a better chance to get their reputability across. In these cases, the increase in general attention would also lead a more information rich environment.

4.1.3 Monopolist with Many Articles

We now relax the assumption that the monopolist can only supply a single article, and instead allow it to supply many ($M \geq 2$). We assume that the number of topics K is large so that $M \leq K$ and the provider is not forced to produce two pieces of content on the same topic if he wants to supply all M articles.[4]

An equilibrium in this setting consists of a sequence of optimal stopping times (per Lemma 4.1.1) for the reader but also probabilities that each of the articles supplied contains information. We let $p^{(i)}$ denote the probability that article i contains information in equilibrium and for convention let these probabilities be ranked in descending order.[5] Recall that by Lemma 4.1.1, the consumer employs stopping times $\{\tau^{(1)}, \tau^{(2)}, \ldots\}$ in equilibrium. An equilibrium can therefore be defined as the vector of tuples $((p^{(1)}, \tau^{(1)}), (p^{(2)}, \tau^{(2)}), \ldots, (p^{(M)}, \tau^{(M)}))$ which

[4] This is to avoid decreasing marginal value from information signals. For example, an unbiased Gaussian signal s_1 about θ_k^* improves the estimate $\hat{\theta}_k$ more than the second signal s_2 does (given s_1 already). This will naturally change the consumer's demand for information when there are more articles, which is tangential to the problem we study.

[5] This ranking is purely for convention of the analysis and not observable to the consumer.

more concisely can be written as $(\boldsymbol{p}, \boldsymbol{\tau})$. Note, that while there were at most two equilibria with a single article (per Theorem 4.1.1), there can be many more equilibria for $M \geq 2$, including equilibria where strictly less than M articles are produced.

Characterization of Information-Rich Equilibria

The following result characterizes the richness of equilibria as a function of M.

Theorem 4.1.2. *For any $M' < M''$, there exists some $\tau^*_{\mathbb{M}} > 0$ such that if $\tau_C < \tau^*_{\mathbb{M}}$, the most information-rich equilibrium with M'' articles is richer than the most information-rich equilibrium with M' articles.*

Under the condition that readers do not spend too much time on catchy articles ($\tau_C < \tau_{\mathbb{M}}$), Theorem 4.1.2 establishes that an increase in the monopolist's technological ability, allowing it to produce and populate more articles, always results in greater information provision (in *some* equilibrium). The intuition is that a monopolist does not want to self-cannibalize by producing many articles of varying nature (some catchy, some information, possibly some both or neither) when it is not competing with other content providers for the user's attention. Given that the consumer has limited time to read, there is greater likelihood the consumer will never get an opportunity to process some articles, making the provider's investment into each additional article less and less profitable.

We first remark that the condition on τ_C cannot be dispensed with, and in fact $\tau^*_{\mathbb{M}}$ will be decreasing as M' and M'' increase. The reason stems from the fact that investments in information, relative to catchiness, decay in value as more articles become available. This occurs precisely because catchy articles are a fixed time investment for readers, but non-catchy articles with uncertain informational value admit shorter and shorter stopping times when more come into existence. This is because with each passing second, the value of a non-catchy article diminishes as it becomes less likely to be one with information; given that there are many more such articles at the consumer's disposable, she is unlikely to spend as much time reading it. As the reader jumps more quickly between non-catchy articles, this in turn leads to fewer quality investments from providers, which further reduces stopping times from readers. Therefore, at some point, providers are better off supplying catchy articles, which stops the spiral and imposes a lower bound on readership. Whether this leads to *more information*, in addition to some catchiness, is generally ambiguous.

Finally, we note that while Theorem 4.1.2 guarantees that increasing the number of available articles at least weakly improves information richness, there are many examples where this is improvement is strict. For instance, when $M = 2$, it will generally be the case that the expected quality of each of the two articles will be less than the expected quality of the single article produced in the $M = 1$ setting. However, overall both articles provide a more holistic view of the state $\boldsymbol{\theta}$ and provide strictly more information compared to the case of only a single article.

Characterization of Information-Poor Equilibria

Our next result shows that while the most information-rich equilibrium is generally richer with more articles (Theorem 4.1.2), there also exist information-poor equilibria that are poorer when more articles are able to be produced.

Proposition 4.1.2. *There exists $M^* > 0$ and $\tilde{\tau} > 0$ such that if $M > M^*$ and $\tau_C < \tilde{\tau}$, all equilibria with one-article capacity are more information rich than the least information-rich equilibrium with M articles.*

Taken together, Theorem 4.1.2 and Proposition 4.1.2 show that more advanced technological capacity can lead to better or worse information depending on the equilibrium, even when there is a sole content provider. Similar to the Coase conjecture, a monopolist who can supply a large number of articles is competing with itself over time for attention. Even if the articles contain information with high probability, the reader has such a large set of articles to choose from that she will continually switch attention until an article provides an immediate reputability signal. This causes all information to unravel in an equilibrium where the full capacity of M articles are produced.

In this information-poor equilibrium, while the content provider would prefer to commit to producing fewer articles, he cannot credibly do this. In an equilibrium with many non-catchy articles, it will never be the provider's best response to supply content that is highly likely to contain information because it will barely be read. If the provider produces articles on the basis of which topics are the easiest to convey information, it suffers a credibility problem: it will always want to masquerade as if it is knowledgeable about all topics, even though such a realization is highly unlikely. As a consequence, the monopolist's ability to produce too many articles in fact inhibits its ability to have any read in any serious capacity in this information-poor equilibrium.

4.1.4 Competition across Providers

Finally, we suppose we have N content providers who compete for user attention. We fix the number of articles M and divide these up equally among N providers.[6] This serves to isolate the main competitive forces of the model instead of simultaneously increasing the amount of content that could be available. Our main result on competition is given next.

Theorem 4.1.3. *There exists $\tau^*_{\mathbb{C}}$ such that for $\tau_C < \tau^*_{\mathbb{C}}$:*

(a) For every equilibrium with $N > 1$ providers, there is an equilibrium with one provider that is more information rich.

(b) For every equilibrium with one provider, there is an equilibrium with N providers that is less information rich.[7]

Theorem 4.1.3 shows that holding the total number of articles fixed, competition typically makes information provision worse. A key observation lies in thinking about the same number of catchy and non-catchy articles with their same likelihoods of containing information under different competition structures. The expected read time will be the same under both, but how this expected reading time is partitioned across providers (and whether it can be supported in equilibrium) will depend on the the structure of competition. Let us consider both directions of the statements in Theorem 4.1.3.

For part (a), take any equilibrium with N providers and we will construct an equilibrium with one provider that is (weakly) more information rich. In the equilibrium with N providers, consider the same collection of articles (with their associated catchiness attributes and likelihoods to be informative). We consider the strategy profile where the monopolist offers the same collection of articles, and we look for profitable deviations. If there are none, we have an equally information rich equilibrium. If there are, it must involve switching articles from catchy to non-catchy or non-catchy to catchy (or both); otherwise, the deviation involves higher or lower likelihood of investing in quality, which is unobservable to the consumer, and thus would have also been a profitable deviation for a provider in the original equilibrium under competition. Only the net switches from catchy to non-catchy and non-catchy to

[6] Implicitly, here, we will assume that M is a multiple of N.

[7] For future work, the goal is to strengthen this result a little. In particular, I would want to compare any two $N \neq N'$ instead of comparing to just a monopolist. But this makes the argument more difficult, so due to a lack of time, I will have to settle for this weaker version.

catchy matter, because these appear as the same to a consumer. If the number of catchy articles increases in a profitable deviation, these must get higher expected reading time (to compensate for the catchiness investment), which means the consumer tends to read them sooner; however, each additional catchy article will cannibalize the others. This implies that it must be a profitable deviation in the original equilibrium for some provider to do a uni-lateral switch to catchy (for a subset of their articles), an obvious contradiction, so the only candidate for a profitable deviation includes switches from catchy to non-catchy. Using the fact this deviation is profitable for a monopolist (by assumption), one can construct an equilibrium with more non-catchy articles but which are more likely to be informative than their catchy counterparts when $\tau_C < \tau_C^*$, establishing the part (a) of Theorem 4.1.3.

The argument is similar for part (b); take any equilibrium with a monopolist and we will construct an equilibrium with N content providers that is more information poor. Using the same key observation as before, we will take the monopolist equilibrium and split up the articles among the N content providers and consider whether this is still an equilibrium. A profitable deviation must take the form of switching some catchy articles to non-catchy or vice-versa, for at least one of the N content providers. If switching some articles to non-catchy is profitable for one provider, it must decrease the reading times for other providers (otherwise it would be a profitable deviation for the monopolist too). Of course, this means catchy articles have lower expected reading times than non-catchy articles, which is a contradiction (why invest in catchiness then?), so profitable deviations can only exist by switching non-catchy articles to catchy ones. As before, this allows one to construct an equilibrium with N providers with more catchy content but which is overall less information rich.

Finally, we remark that competition can be *beneficial* for larger values of τ_C. In the archetypal setting of Theorem 4.1.3, catchiness and information are substitutive, so competition that tends to foster catchiness as a way to attract users is detrimental to information richness. However, for higher values of τ_C, information and catchiness can act as complementary; instead, competitors might push each other to simultaneously compete on article appeal *and* information. This is perhaps counterintuitive but in the same vein as Proposition 4.1.2: catchiness can prevent an unraveling of information provision when there are far too many articles for the limited attention of the consumer, as it requires a certain read time to absorb the full entertainment value.

4.2 An Experiment on Algorithmic Ranking: User Behavior, Platform Incentives, and Policy

> "Quantifying the impacts of algorithmic ranking is quite difficult, even with access to proprietary data. This is not only because of the complexity of these technical systems, but due to people's complex and often strategic responses to changes in algorithms. We lack clear evidence about broader benefits or harms of algorithmic ranking."
>
> — Dean Eckles (2021)
> Testimony before Senate

In this section, we lay out a proposal and present a set of partial results for an experiment intended to capture the impacts from algorithmic ranking.[8] While an oversimplification of platform algorithms, the goal of this experiment is to implement intuitive ranking algorithms on a news feed platform such as Facebook and Twitter, and to assess the impacts on user engagement. We are especially interested in how participants engage (in the form of "likes" or "shares") with content such as misinformation (as discussed in Chapters 2-3) and sensational or catchy content (as discussed in Section 4.1), and whether they spend more time lingering on the feed as a result of the ranking algorithm.

4.2.1 Mission Statement

Background and Motivation. In order to boost engagement, platforms develop recommendation algorithms that strive to maximize the time users spend on their site. We aim to understand these recommendations and their potential consequences. For a basic illustration, consider three motivating questions:

1. How do users make decisions about the content they engage with?

2. How do platform recommendations account for the decision making of users in (1) to make them spend the most time on their platform?

[8] I would like to especially acknowledge my co-author, Charles Lyu, for helping substantially with drafting this section of the thesis.

3. How do users respond to the recommendations of the platform in (2)?

All three questions encompass highly complex ideas and often non-measurable interactions from both users and platforms, while intertwining with each other. However, they serve as starting points for our proposal. By experimentally analyzing the impacts of certain recommendation methods on social media users, particularly users' engagement patterns towards these different ranking algorithms, we aim to provide some insights on what could reasonably happen in real-world setups, especially (1) and (2).

For example, users may engage most with political content that agrees with their ideological belief (question (1)). Given this behavior, the platform may want to only recommend content that is ideologically congruent to the users' political beliefs (question (2)), regardless of whether this content is reliable (e.g., potentially likely to contain misinformation). This can be bad from a societal perspective for two reasons. First, it can lead to a one-sided view of the world (an "echo chamber" of similar content circulating) that neglects a holistic picture of all opinions. Second, it creates an ideological filter that deprioritizes reputable content to push more ideologically similar content, leading to overall less accurate content being recommended.

But finally, given the priorities identified by question (2), users may respond in unexpected ways. For example, users may more aggressively block (or "unfriend") others of opposing ideology once algorithmic ranking (AR) is introduced. Does this amplify echo chambers or dampen them? The simplistic observation is that it should amplify them. AR removes "friends" outside the ideological echo chamber and makes content more homogenous within one's sharing network circle. However, there is a more subtle effect. Because AR avoids bombardment of ideologically opposed content, it leads to less likely "unfriending" of those with divergent beliefs. Overall, the impact on diversity and well-roundness of the content that appears in the news feed is unclear.

Objective and Focal Questions. Here are the two main goals for the experiment with a small number of details.

- How do different methods of algorithmic ranking affect user engagement of the news feed, including items with different political leanings?
  - Examples of specific questions: How does each ranking method affect the engagement

of both prioritized and unprioritized content, compared to randomized news feeds? Does the effect differ for ideologically congruent vs incongruent items? Do recommendations based on both friendship and preferences have greater impacts than the individual methods combined?

- How do different methods of algorithmic ranking affect engagement with misinformation?

Admittedly, some or all of these questions may be better examined using other experimental or empirical methods, instead of our proposed lab experiment with Qualtrics. While we are indeed looking into possible alternatives, our current impression is that the Qualtrics experiment is still the best.

4.2.2 Experimental Procedure and Design Choices

Overview. Our main experiment consists of two waves, Wave 1 and Wave 2.

Wave 1 is a small-scale survey aimed at creating artificial social media profiles to be presented during Wave 2. Each Wave 1 Participant (W1P) contributes to one such profile based on their demographic information, interests, opinions on sociopolitical issues, interaction with a randomized news feed, and preferences for advertisements.

Wave 2 is the crux of the study, where we analyze user behavior to answer our focal questions. Each Wave 2 Participant (W2P) first chooses to follow one or more profiles of W1Ps as "friends". They will then be presented with two news feeds: a fully randomized feed, and an algorithmically ranked feed that prioritizes articles based on friend interaction, their own preferences, or both. Their behavior across the two news feeds will be compared to examine the impacts of ranking on user engagement.

News Headlines and Pretest. Prior to the two waves, we conducted a pretest to aid in our selection of news headlines to be displayed in news feeds. The pretest gathers crowdsourced data on each headline's topic and characteristics.

Initial Selection of Headlines. We collected 339 news headlines in June 2021 that were used for the pretest, before selecting 230 of them for the main experiments. These headlines cover the following topics: Politics, US and local news, World news, Economy and business, Science and technology, Health and COVID-19, Sport and entertainment. Each headline was given a preliminary classification into one of these topics based on our own judgment. Roughly half of the headlines were political news.

62 headlines contain misinformation. Most of them were from articles that were classified as false by professional fact checkers, such as Snopes and PolitiFact. This approach closely follows the guide in Pennycook et al. (2021a). A small number of these headlines were taken directly from websites with a known history of publishing "fake news" articles, such as the Gateway Pundit. We chose misinformation articles related to the US politics and COVID-19.

The remaining 277 truthful headlines were collected from the mainstream media in the United States. A large number of them were from the following outlets: CNN, Fox News, the New York Times, the Hill, Wall Street Journal, NBC News, Washington Post and the Guardian.

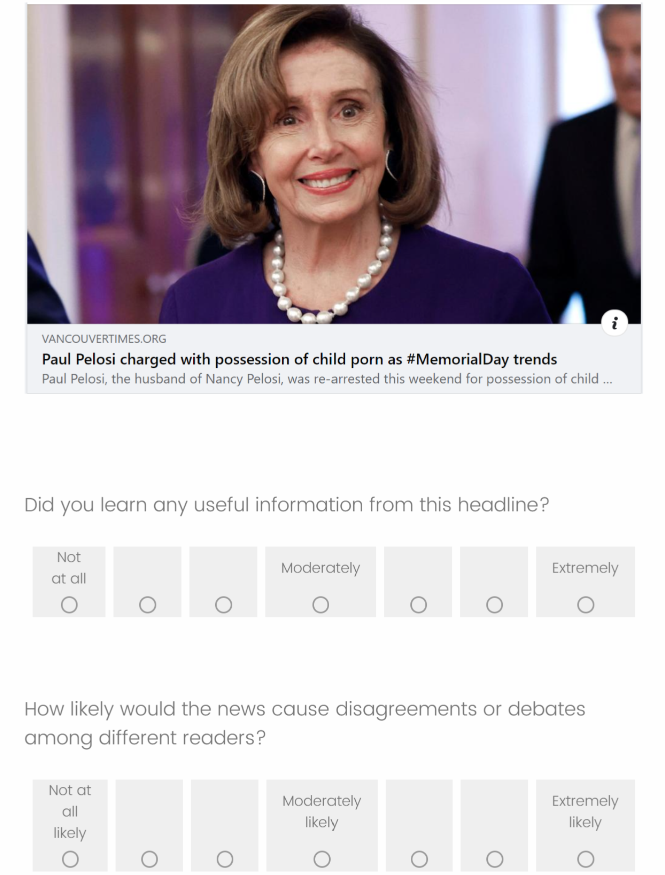
Most truthful news articles were gathered in June 2021, while most misinformation articles were fact checked between March and June. A notable number of political articles focus on ongoing events during this time frame, such as Roe v Wade, the January 6 hearings, and school shootings throughout the United States in May. A breakdown of number of headlines by topic can be found in Table 4.1.

Pretest Design and Questions. A pretest was conducted on Amazon Mechanical Turk. The preliminary goal is to gather crowd-sourced assessments of each headline, which have two main uses: (1) for selection of a balanced subset of news headlines for the main experiments, and splitting it into two pools; (2) to be used as features for preference-based recommendations.

Each participant is presented with 10 random headlines, and for each headline, they are asked nine questions from the following categories:

1. How well does this headline fit into each of the seven news topics?
2. How well does this headline score on each of the five news characteristics (detailed below)?
3. How favorable would this headline be to Democrats and Republicans, respectively (assuming it is entirely accurate)?
4. If you were to see the above article on social media, how likely would you be to "like" and share it?

Questions from categories 2 and 4 use 7-point unipolar scales from “not at all” to “extremely”. Category 1 uses 5-point unipolar scales, while category 3 uses 7-point bipolar scales from “extremely unfavorable” to “extremely favorable”. The order of choices for all questions is randomly flipped for half of the participants. Aside from the headline-specific questions, the survey also contains three attention check questions, and a number of demographic questions at the end. Figure 4-2 shows the user interface of the pretest survey.



VANCOUVERTIMES.ORG

Paul Pelosi charged with possession of child porn as #MemorialDay trends

Paul Pelosi, the husband of Nancy Pelosi, was re-arrested this weekend for possession of child ...

Did you learn any useful information from this headline?

Not at all			Moderately			Extremely
○	○	○	○	○	○	○

How likely would the news cause disagreements or debates among different readers?

Not at all likely			Moderately likely			Extremely likely
○	○	○	○	○	○	○

Figure 4-2. User interface of the pretest survey. More questions on the same headline are available at the bottom and on a subsequent page.

News charactistics. We consider the following characteristics for each news headline, which provide quantitative measurements of a news article from different dimensions.

- **Veracity**: Does this article may contain misinformation or false information? How likely is this headline true?
- **Conflict**: Does this headline depict a controversy, conflict or dispute? How controversial is the event or opinion described in the headline?

- **Human interest**: How entertaining or funny is this headline?

- **Surprise**: How surprising or unexpected is this headline?

- **Relevance**: How important is this headline? Does it offer relevant and useful information to readers who are interested in this topic?

The chosen characteristics were derived from literature on journalism studies. We specifically consider news values that (1) are prevalent among news articles, especially popular ones on social media; and (2) can possibly demonstrate enough variety among these articles for better user profiling.

A major motivation for considering several characteristics (instead of just veracity) is to design a simple preference-based recommendation algorithm that considers these aspects, along with news topics and political polarity, when suggesting headlines to users. For example, the algorithm may detect from earlier interactions that a certain user prefer local news that are favorable to Republicans, controversial, and surprising. The algorithm may then suggest news articles with one or more of these attributes.

The survey contains three to four questions for each news characteristic with different phrasing. This is mostly because some characteristics can have slightly different interpretations; a few questions are also reverse-coded (e.g. how boring instead of how funny) to be more robust to inattentive survey participants. Each participant is presented with one randomly chosen question for each news characteristic, and the five chosen questions are also shuffled.

Attention checks. The survey contains three attention check questions: one at the start of the survey with question “Puppy is to dog as kitten is to”; one after the fifth headline asking the participant to recall three keywords from any headline they have read so far; and one after the final headline with question “What is your favorite color?”, but with a prompt that instructs participants to choose red and green, regardless of their favorite colors. Participants who fail the first attention check were automatically deemed ineligible and redirected to the end of the survey. All responses to the other two attention check questions were reviewed manually; those who did not provide genuine responses were not given compensation for completing the survey, and their responses were excluded from data analysis.

Processing of Headlines from Pretest Results

After pretest results were gathered and responses that failed any of the attention checks were removed, the data was processed via the following steps:

Compute average scores. For each headline, scores for each news topic, each news characteristic, favorability to each political party, and likelihood of liking and sharing were computed from average responses of participants who were assigned this headline. Nine headlines were removed from consideration as they received an insufficient number of responses (11 or fewer).

Reassign topic classifications. Each headline was recategorized into one of the seven news topics based on user responses. Typically, the topic with the highest average score was chosen as the new topic label; unless some other topics had a score within 0.2 of the maximum, in which case a decision was made manually. While most headlines retained their initial topic classification that we picked, some headlines received a different classification.

Compute political polarity scores. The pretest contains two questions on how favorable the headline is to Democrats and Republicans respectively. The two distinct favorability scores can be useful in preference-based recommendations, and they reveal more details than a single score (for example, some articles may be favorable to Democrats while not necessarily hurting Republicans, which are different from articles unfavorable to Republicans). Nevertheless, a single score on political leaning is often more convenient for data processing. Thus, for headline i that belongs to the topics Politics or Local News, we define a polarity score p_i as the following:

$$\begin{aligned} p_i &= z_{i,R} - z_{i,D}, \\ \text{where} \quad z_{i,R} &= \frac{x_{i,R} - \mu_R}{\sigma_R}, \\ z_{i,D} &= \frac{x_{i,D} - \mu_D}{\sigma_D}. \end{aligned}$$

Here, $x_{i,R}$ is the Republican favorability score of headline i (average of all responses); the higher the $x_{i,R}$, the more favorable headline i is to Republicans. μ_R and σ_R are the mean and standard deviation of the $x_{j,R}$ scores of all headlines j in the same topic as i, and $z_{i,R}$ is the z-score of headline i's Republican favorability. $x_{i,D}$, μ_D, σ_D and $z_{i,D}$ are defined similarly on Democrat favorability scores. Therefore, a positive p_i indicates headline i is generally more favorable to Republicans than Democrats, while a negative p_i means it is more favorable to Democrats than Republicans.

The main reason of defining p_i using z-scores instead of raw scores is to resolve the bias in pretest responses, as a significantly greater number of survey participants self-identify as Democrats than all other options (Republican, independent, others).

Note that polarity scores are defined separately on political news and local news, with slightly different means and standard deviations. This is because the two subsets of headlines have different underlying distributions: for example, a greater number of local news headlines are politically neutral compared to political headlines. For the same reason, polarity scores are not computed for headlines with a final topic classification that is not political or local news, as most of these headlines are intended to be neutral to both parties.

Subsampling of Headlines for the Main Experiments

After we obtain crowd-sourced measurements of each headline, a final subset of headlines is selected to be used for the main experiments (Wave 1 and Wave 2). This subset is further divided into two equal pools: Pool A is used for the random news feed in Wave 1 and the algorithmically ranked news feed in Wave 2, while Pool B is used for the random news feed in Wave 2. This decision is elaborated in Section 4.2.2.

Pretest data is important in selection of the two pools to ensure the chosen headlines are roughly balanced in political polarity on both extremes (i.e. left-leaning headlines favor Democrats just as much as right-leaning headlines favor Republicans), and that Pools A and B are similar in both quantities of headlines and average scores of important attributes.

Table 4.1 shows a breakdown of headlines chosen for each pool by topic. Headlines are chosen by rejection sampling, with the following constraints:

1. For each topic, the two pools should have similar average scores in like intentions, share intentions, and veracity (perceived by pretest participants). Other news characteristics should preferably have similar scores, but not as a hard constraint.

2. For the topics Politics and Local News, in addition to constraint (1):

Topic	Politics	Local	World	E&B	Tech	Health	S&E	Total
Initial topic	131	48	25	26	32	49	28	339
Misinformation	45	0	0	0	1	15	1	62
Crowd-sourced topic	84	90	26	34	28	41	27	330
Misinformation	22	21	2	2	2	10	1	60
Waves 1 & 2	60	60	20	20	20	30	20	230
Misinformation	14	14	0	0	0	6	0	34
Waves 1 & 2 per pool	30	30	10	10	10	15	10	115
Misinformation per pool	7	7	0	0	0	3	0	17

Table 4.1. Number of pretest headlines for each topic, with our initial manual classification of topics, final classification from crowd-sourced data, and the subset selected for the main experiment (Waves 1 & 2). "Crowd-sourced topic" excludes headlines with insufficient pretest responses. Totals include misinformation headlines. Local = US and Local News; E&B = Economy and Business; Tech = Science and Technology; S&E = Sports and Entertainment.

- Conditioned on true, fake or all headlines, the two pools should be similar in veracity, political polarity, and the number of left-leaning, neutral, and right-leaning headlines.
- Conditioned on true, fake or all headlines, the average polarity of left-leaning headlines and that of right-leaning headlines should be similar in magnitude.
- Taking any combination of true, fake or all headlines, *and* left-leaning, neutral or right-leaning headlines, such headlines in the two pools should still be similar in veracity and polarity.

3. For the topic Health, in addition to constraint (1), the two pools should also have similar average veracity scores when conditioned on true headlines or fake headlines.

The additional constraints on political and local news are to ensure balance in true vs misinformation headlines, and in political polarization of headlines. Since there is also a large number of COVID-19 misinformation headlines in the dataset, the additional constraints on misinformation are also applied to health. Categorization of left-leaning, neutral and right-leaning headlines is done with a threshold on polarity scores.

Wave 1

Purpose of Wave 1

While results of Wave 1 are not central to answering our focal questions on the effects of algorithmic ranking, this initial phase is necessary for the following reasons:

- Each response from a Wave 1 Participant (W1P) will be used to create one artificial social media profile. These profiles will be offered to Wave 2 Participants (W2Ps) to be added as friends, upon which W2Ps may be recommended news headlines that were liked and shared by their friends during Wave 1.

- We also use Wave 1 to build a demographic baseline on preferences of difference brands of certain products, such as shoes, snacks and vacation spots. This is for an ad-based measure of attention to be deployed in Wave 2.

- Wave 1 also serves as a sanity check on the effectiveness of the experimental platform, the random news feed and how users interact with it.

Survey Flow

The survey has four main components. The first four are conducted on Qualtrics, after which participants will be redirected to Yourfeed, an experimental platform for behavioral research on social media, where they will interact with a random news feed as the final component.

Personal information. Each W1P will first answer some basic questions about their demographics, hobbies, and political beliefs. Items from all sections may be shown on the "user profiles" that will be presented during Wave 2.

Essay writing. Each W1P will be instructed to write two essays, one about themselves, and one on a sociopolitical issue. The prompts will be randomly chosen from a pool of 10 questions, 5 political and 5 apolitical. They are expected to write two to three sentences, with a minimum of 150 characters, as if they are replying to a new friend.

We will filter out fraudulent and low-quality responses, both manually and by running plagiarism checks.

Brand preferences. Each W1P will be presented with several categories of items (shoes, snacks, vacation spots, etc), and several brand names for each (e.g. Nike, Adidas, Puma). For each category, they will choose the brand they're most likely to buy.

This phase is to establish a demographic baseline of preferences, i.e. the proportion of the general population who would choose each brand. The data will be used for comparisons during Wave 2 (where W2Ps may be presented with ads) in order to measure attention spent on the news feed. More details are discussed in Section 4.2.2.

Random news feed. Each W1P is presented with a random scrolling news feed of news headlines on Yourfeed. A random subset of 50 headlines from Pool A are displayed in uniformly random order. Participants can like or share any headline, and their dwell time on each headline is also measured.

The articles that each W1P liked and shared may be displayed to Wave 2 Participants who choose to add this W1P as a friend.

Questions and Essay Prompts

Questions on personal information:

- Demographics: Participants will be asked about their age, gender, state of residence, education, income and ethnicity. The first three may be presented in user profiles later.
- Hobbies: Participants will choose at least one from: Games, sports, outdoor recreation, movies & TV series, art, music, collecting, reading, travelling, social activities, and others. For each one that they choose, they can optionally elaborate with a list of specific activities (e.g. basketball, Call of Duty, stamp collection). All hobbies they provided will be presented in user profiles.
- Political beliefs: Participants will provide their political preference, views on social issues, and views on economic issues. Their political preference may be presented in user profiles.

Political essay prompts:

- Are you satisfied with Biden's term so far? Why or why not?
- How do you feel about the overturning of Roe v Wade? Do you support abortion rights, and why or why not?
- Do you support wealth redistribution? Why or why not? What do you think are ideal approaches to income inequality in the US?
- Which party do you think will win the 2024 Presidential Election? Explain why. You may also mention any candidates that you think have the highest chance of winning.
- What are your thoughts on the Jan 6 capitol riots and the hearings? How much do you think Trump was responsible for this, and why or why not?

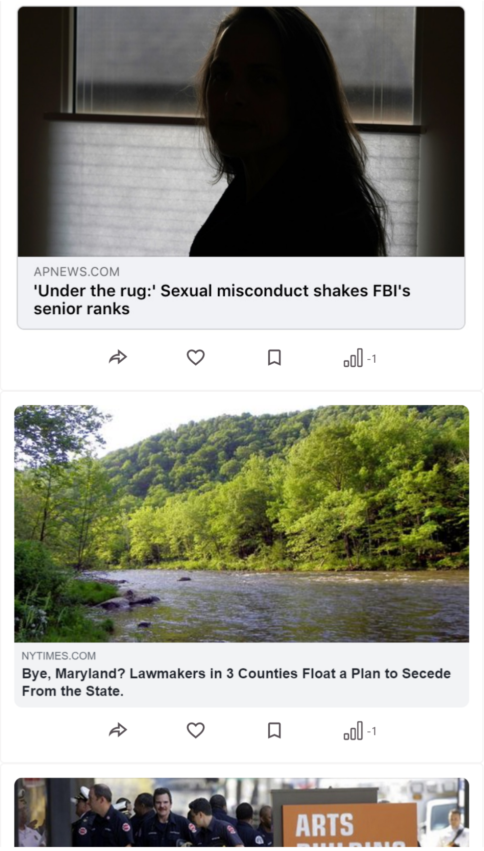
Apolitical essay prompts:

- What are some things you do that you think nobody else would?
- If you win the $1 billion lottery jackpot, what would you do?
- What were your childhood dreams and wishes? Have you accomplished any of them so far?
- Name 3 items on your life checklist that you have not yet achieved. How do you plan to check them off?
- What are the top 5 things that make you happy in life?

Yourfeed Interface and Integration

Yourfeed is an experimental research tool for conducting behavioral research on social media (see Epstein and Lin (2022) and Epstein et al. (2022)). The platform offers an interface that resembles a news feed on a social media platform (Figure 4-3), where participants can scroll through a list of news headlines with thumbnails vertically. Each news item is accompanied with a "like" and "share" button (and optionally with the number of likes and shares, which we do not need for our experiment).

There are two reasons why we choose Yourfeed as the platform for all of our news feed interactions in Waves 1 and 2. The close resemblance of its interface to real social media



APNEWS.COM

'Under the rug:' Sexual misconduct shakes FBI's senior ranks

NYTIMES.COM

Bye, Maryland? Lawmakers in 3 Counties Float a Plan to Secede From the State.

Figure 4-3. User interface of Yourfeed

platforms makes user behavior on the platform better approximations to real-world data, as compared to traditional approaches using Qualtrics. Indeed, Epstein and Lin (2022) and Epstein et al. (2022) show there are significant differences in user behavior between studies conducted on Yourfeed vs Qualtrics, where participants are less likely to like and share articles on Yourfeed, likely closer to reality. Furthermore, Yourfeed provides measurement of dwell times, or the duration each participant spends on each news item. This provides a more complete view of user behavior on social media platforms, in addition to liking and sharing behavior that were typically the focus of prior studies.

Wave 2

Survey Flow Overview

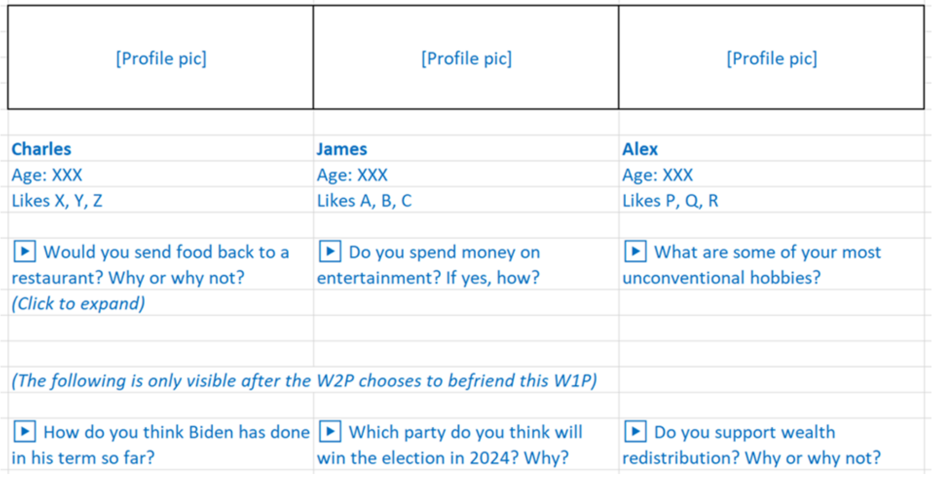
During the main experiment (Wave 2), participants will form “friendships” among some other participants (Phase I), interact with a random scrolling news feed (Phase II), and then interact with an algorithmically sorted news feed that prioritizes articles shared by friends, or articles with similar characteristics as the user's prior selections, or both (Phase III).

Wave 2 Phase I: Friendship Formation

Each Wave 2 Participant (W2P) will first be presented with three apolitical essay prompts and three political essay prompts, randomly chosen from the pool of five each that were given to W1Ps. They will be instructed to choose at least two headlines out of the six that they are interested in, but are free to choose more. There are no restrictions on the prompts they choose.

We allow W2Ps to choose their own essay prompts, in order to examine if users prefer to choose their friends based on political alignment more than non-political interests, and if doing so may amplify the filter bubble effect.

For each chosen prompt, the W2P will then be given the profiles of 8 W1Ps who responded to that prompt, for a total of 16 or more profiles in random order. Each profile has a fake name and an avatar, as well as demographic information and hobbies from real Wave 1 responses. Each profile will be accompanied by the chosen essay prompt, which the W2P can expand to read the essays. The W2P be told that these are participants who have completed the survey previously, and that they might be interested in knowing them and adding them as friends. Figure 4-4 gives a rough sketch of how the W1P profiles are displayed.



[Profile pic]	[Profile pic]	[Profile pic]
Charles	**James**	**Alex**
Age: XXX	Age: XXX	Age: XXX
Likes X, Y, Z	Likes A, B, C	Likes P, Q, R
▶ Would you send food back to a restaurant? Why or why not? *(Click to expand)*	▶ Do you spend money on entertainment? If yes, how?	▶ What are some of your most unconventional hobbies?
(The following is only visible after the W2P chooses to befriend this W1P)		
▶ How do you think Biden has done in his term so far?	▶ Which party do you think will win the election in 2024? Why?	▶ Do you support wealth redistribution? Why or why not?

Figure 4-4. Sketch of the page on the Wave 2 survey where users can see profiles of W1Ps and choose to add them as friends

When the W2P expands an essay, they have the option to add that W1P as a friend. They will need to add at least three W1Ps as friends. Once they finish all friend selections and proceed, the W2P needs to pick one of the W1Ps they added, and write a short response to explain their choice. (Example prompts: Why did you choose this person as a friend? Why do you want to have lunch with them? How would you respond to what they wrote if you were chatting with them in real life?)

The W2P will then see the other essay that each W1P friend has written (the prompt that was not chosen by the W2P). They will give a rating (1 to 5) for each W1P in regards to how their impression on this W1P has changed, and write a short response to justify their rating for the particular friend they wrote a paragraph on earlier. All the chosen F1Ps will become “friends” of the W2P for later phases, regardless of their indications after the second essay.

All written responses from W2Ps will not be used at all, but these exercises aim to improve W2P’s engagement with the experiment. After the experiment, we will filter out fraudulent and low-quality responses, both manually and by running plagiarism checks.

Wave 2 Phase II: Random News Feed

Each W2P is presented with a random scrolling news feed of headlines on Yourfeed. A random subset of 50 headlines from Pool B are displayed in uniformly random order. W2Ps can like and share any headlines.

The main goal of this phase is to build a preference profile for each W2P based on the types of articles they liked, shared, and spent more time reading. This will be the primary basis for preference-based recommendations in Phase III.

Wave 2 Phase III: Ranked News Feed and Treatment Groups

Each W2P is randomly assigned to one of the following four treatment groups. They are presented with a strategically sorted news feed, showing a subset of 50 headlines from Pool A, chosen based on the treatment group.

(a) Control group: 50 random articles are chosen and shown in uniformly random order.

(b) Friendship group: Articles that were previously shared or liked by the user’s "W1P friends" are prioritized.

(c) Preference group: Articles with similar topics, characteristics and political leaning to the user’s prior selections in Phase II (and their self-reported political position) are prioritized.

(d) Friendship and preference group: Articles based on both groups (b and (c are recommended in some combination.

Regardless of the treatment group, all articles shared by W1P friends will be accompanied with a tag like “Person X shared this”. But groups other than (b and (d will likely get them much later in the news feed, if at all.

Prioritization of articles. Depending on the treatment group, a list of prioritized articles from Pool A is generated in a particular order. For group (b, shared articles appear earlier in this list than liked articles. For group (c, articles are sorted by a similarity score based on news topics and characteristics, with the most similar article appearing first.

Each W2P is assigned a random variable between 0 and 1, which determines the “degree of prioritization”, specifically the probability of each item in the feed (50 articles, top to bottom) being pulled from the list of prioritized articles. For example, if the random variable is 0.7, the first item in the news feed has 70% chance to be the first article in the priority list, and 30% chance to be a random article from the entire Pool A. The chosen article will be removed from the pool and the list. This procedure is repeated until a full list of 50 articles is generated and presented to the W2P.

Knowledge of news feed being ranked. In each treatment group, half of the participants are not informed of the feed being algorithmically ranked, and the other half will see a prompt “Your news feed has been ranked based on your preferences and/or what your friends have shared”. This also applies to the control group (a, where half of them get a uniform random feed with no prompts, and half see the prompt but there is actually no ranking in place. We will mention in the debrief for these participants that their feed was actually fully randomized.

Measuring Attention via Ads

In addition to measuring attention directly using dwell time, we present a novel approach of an indirect measure via ads in the news feed. Each time a W2P interacts with an article in Phase III, an ad will show up roughly 2-4 places after this item. The ad will be from a randomly chosen category of items (shoes, vacation spots etc), but for each category, each W2P will always get a fixed brand. For example, a W2P may get 3 shoe ads and 2 vacation ads, but all shoe ads will be

Nike and all vacation ads will be Mexico. Another W2P may get all Adidas ads for shoes and all Japan ads for vacation.

At the end of Phase III, the W2P will choose the brand they're most likely to buy for each category of items. The collective responses of all W2Ps in each experimental group will be used to compare against the demographic baseline obtained from the Wave 1. Our hypothesis is that W2Ps will show an increase in preference for the brand that appeared in their ads (e.g. 60% of W1Ps prefer Nike, but the W2Ps who were recommended Nike show 65% instead).

The increase in preference is a proxy for attention spent on the news articles (and thus attention to ads), in addition to dwell time measured directly on Yourfeed. While more noisy than dwell time, this measure of attention aligns well with the goal of profit-maximizing social media platforms such as Facebook. While time and attention spent on the news feed are important to them, their ultimate goal is to maximize ad revenue. Here, we're measuring "ad revenue" directly, so the results may be more applicable in real life settings.

Use of Two Different Pools of News Headlines

As noted in previous sections, after we selected the set of news headlines to be used in the main experiments (Waves 1 and 2), we divided them further into two subsets, Pool A and Pool B. This design decision arose from two needs of our experiment. For one, W2Ps need to interact with both a random news feed (Wave 2 Phase II) and a ranked news feed (Wave 2 Phase III), so that we can derive their preferences from Phase II, as well as comparing their behavior in these phases to study the impacts of ranking algorithms. For another, articles liked and shared by W1P friends need to be highlighted in the ranked news feed, and prioritized for treatment groups (b and (d.

The second need means Wave 2 Phase III must use the same pool of news headlines as Wave 1, while the first need means a different pool must be used for Wave 2 Phase II.[9] In practice, we assign Pool A to Wave 1 and Wave 2 Phase III, and Pool B to Wave 2 Phase II. We also enforced constraints to ensure the two pools are similar in important attributes, as discussed in Section 4.2.2, to set up a valid comparison of user behavior towards the two news feeds.

[9] While it is possible for both Phases II and III to draw from a single pool of headlines, articles liked and shared by W1Ps may have already appeared in the random news feed, and thus can't be shown again in the ranked feed. This may result in not enough articles to prioritize.

4.2.3 Existing Results and Conjectures

Pretest

We received 1,161 responses to our pretest survey, and 686 of them were complete and passed all attention checks. 199 responses were incomplete or completed too late after our data collection phase was already over. 276 responses failed at least one attention check. A detailed breakdown of invalid responses is available in Section 4.2.3.

The 686 valid responses gave ratings to 6,809 headlines[10]. 84 of them were removed after eliminating headlines with 11 or fewer ratings. This section shows results from the remaining 6,725 headline ratings, unless otherwise stated.

Summary

On average, each headline in the pretest received 20.38 ratings. 12.98 of them were from participants self-identified as Democrats, 5.61 from Republicans, 1.72 from Independents, and 0.06 from others.

Table 4.2 shows the number of headlines in each topic and each subset of headlines (all pretest headlines, subset of headlines selected for the main experiments, and each of the two pools), their average favorability to Democrats and Republicans, and polarity scores for political and local headlines.[11] The crowd-sourced favorability scores show a bias towards Democrats: the average Democrat favorability score is 4.89, higher than the average Republican favorability score of 4.71. This is true for each individual news topic, including apolitical topics such as Science & Technology and Sports & Entertainment. We suspect this is due to the majority of participants from Mechanical Turk self-identifying as Democrats. As such, our polarity measure was designed to eliminate the bias.

Table 4.3 shows the averages scores for each news characteristic, like intentions and share intentions. Note that the topics Politics, Local and Health contain a significant number of misinformation headlines, which affect their veracity, like and share scores, as pretest participants indicate they are generally less likely to like or share headlines they consider less

[10] Although we intended each participant to rate 10 headlines, a small number of them received fewer than 10 distinct headlines due to technical difficulties.

[11] As mentioned previously, polarity scores are computed separately for political and local headlines, and not computed for any other topic. The average of 0.00 among the pretest headlines is by design, as polarity scores are defined using z-scores which have zero bias.

likely to be true.

Attribute	Headlines	All	Politics	Local	World	E&B	Tech	Health	S&E
# Headlines	Pretest	330	84	90	26	34	28	41	27
	Main	230	60	60	20	20	20	30	20
	Pool A	115	30	30	10	10	10	15	10
	Pool B	115	30	30	10	10	10	15	10
Democrat Favorability	Pretest	4.89	4.92	4.89	4.86	4.84	4.92	4.88	4.92
	Main	4.88	4.92	4.89	4.86	4.70	4.87	4.93	4.92
	Pool A	4.92	4.95	4.92	4.84	4.77	4.82	4.94	5.09
	Pool B	4.85	4.88	4.86	4.88	4.64	4.91	4.91	4.77
Republican Favorability	Pretest	4.71	4.74	4.66	4.69	4.75	4.62	4.79	4.66
	Main	4.70	4.72	4.70	4.64	4.70	4.59	4.77	4.70
	Pool A	4.75	4.75	4.73	4.68	4.69	4.74	4.85	4.86
	Pool B	4.65	4.70	4.68	4.60	4.71	4.45	4.70	4.54
Polarity	Pretest	N/A	0.00	0.00	N/A	N/A	N/A	N/A	N/A
	Main	N/A	-0.01	0.11	N/A	N/A	N/A	N/A	N/A
	Pool A	N/A	-0.02	0.11	N/A	N/A	N/A	N/A	N/A
	Pool B	N/A	0.00	0.11	N/A	N/A	N/A	N/A	N/A

Table 4.2. Number of headlines and average scores for political attributes. All scores are on a 7-point scale, with minimum 1 and maximum 7. Local = US and Local News; E&B = Economy and Business; Tech = Science and Technology; S&E = Sports and Entertainment.

Examples of Headlines and Responses

Topic classification. Table 4.4 shows examples of headlines from each of the seven news topics, as well as headlines whose topic we picked is different the topic with the highest average score given by pretest participants (thus demonstrating the need for a pretest). Of the 330 headlines, 217 retained their original topic as agreed by participants; 98 had their topics changed[12]; 15 had participants voted on a most popular topic that was different than we picked, but upon manual review, we decided to retain the original topic as its score is close to the top pick.

User beliefs on accuracy and veracity. Table 4.5 shows examples of true and misinformation headlines with some of the highest ans lowest veracity scores from users. This includes headlines where pretest participants most accurately predicted their veracity, and headlines where their judgment is the furthest from the truth. It should be noted that while the vast

[12] A large number of these are misinformation headlines. For convenience, we labeled all of them are political news as long as they were verified to contain misinformation. However, many of them were more appropriately labeled as local news or health (COVID-19 misinformation).

Attribute	Headlines	All	Politics	Local	World	E&B	Tech	Health	S&E
Veracity	Pretest	4.45	4.34	4.45	4.52	4.65	4.60	4.42	4.43
	Main	4.49	4.39	4.43	4.56	4.69	4.65	4.50	4.49
	Pool A	4.49	4.42	4.41	4.55	4.70	4.66	4.50	4.49
	Pool B	4.48	4.36	4.45	4.56	4.69	4.64	4.49	4.48
Conflict	Pretest	4.70	4.92	4.84	4.68	4.33	4.35	4.75	4.29
	Main	4.68	4.94	4.85	4.66	4.22	4.33	4.67	4.26
	Pool A	4.70	4.95	4.88	4.65	4.24	4.34	4.67	4.28
	Pool B	4.67	4.93	4.82	4.66	4.19	4.32	4.68	4.24
Interest	Pretest	4.23	4.17	4.26	4.14	4.24	4.34	4.21	4.23
	Main	4.20	4.11	4.26	4.15	4.11	4.37	4.21	4.18
	Pool A	4.22	4.18	4.26	4.17	4.08	4.41	4.21	4.23
	Pool B	4.17	4.04	4.26	4.14	4.13	4.33	4.21	4.14
Surprise	Pretest	3.99	3.88	4.10	3.91	3.96	4.15	4.01	3.93
	Main	3.95	3.81	4.10	3.91	3.85	4.05	3.99	3.93
	Pool A	3.93	3.78	4.03	3.93	3.83	4.10	3.98	3.98
	Pool B	3.97	3.84	4.16	3.89	3.88	4.00	4.01	3.88
Relevance	Pretest	4.78	4.69	4.84	4.91	4.80	4.76	4.99	4.47
	Main	4.80	4.73	4.83	4.92	4.77	4.74	5.01	4.55
	Pool A	4.81	4.77	4.84	4.95	4.79	4.74	5.01	4.54
	Pool B	4.78	4.69	4.82	4.89	4.74	4.75	5.01	4.56
Like	Pretest	4.29	4.18	4.27	4.33	4.37	4.48	4.28	4.41
	Main	4.27	4.16	4.29	4.35	4.16	4.36	4.26	4.48
	Pool A	4.27	4.17	4.29	4.35	4.15	4.37	4.25	4.48
	Pool B	4.27	4.16	4.28	4.36	4.17	4.36	4.26	4.47
Share	Pretest	4.28	4.14	4.33	4.29	4.32	4.43	4.32	4.31
	Main	4.27	4.17	4.35	4.26	4.11	4.33	4.29	4.39
	Pool A	4.27	4.18	4.35	4.26	4.12	4.32	4.28	4.39
	Pool B	4.27	4.16	4.36	4.26	4.11	4.33	4.30	4.39

Table 4.3. Average scores for apolitical attributes. All scores are on a 7-point scale, with minimum 1 and maximum 7. Local = US and Local News; E&B = Economy and Business; Tech = Science and Technology; S&E = Sports and Entertainment.

majority of our misinformation headlines were verified false by professional fact checkers, some of them could be unintentional or a mistake, instead of intentionally spreading "fake news" for political agenda. This appears to be the case for most misinformation headlines that participants misjudged as true.

Scores. Table 4.6 shows headlines with the highest and lowest average scores on favorability to Democrats and Republicans, the five news characteristics (veracity, conflict, interest, surprise, relevance), and likelihood of liking and sharing the article on social media. Figure 4-5 shows distributions of pretest user ratings for each topic and each attribute for two headlines.

Headline	Initial topic	Final topic
Headlines where participants’ ratings agreed with our manual labels		
Fox News Poll: Record-high negative views of Biden	Politics	Politics
U.S. to propose rule to limit nicotine levels in cigarettes	US & Local News	US & Local News
European leaders visit Kyiv for talks on weapons, EU membership; UK announces new sanctions on Russia	World News	World News
Tesla raises model prices amid global supply-chain shortage	Economy & Business	Economy & Business
Google sidelines engineer who claims its A.I. is sentient	Science & Technology	Science & Technology
Omicron COVID-19 variant gives little immunity boost to those infected: researchers	Health & COVID-19	Health & COVID-19
How a tennis nerd gave Serena and Venus Williams a new lease on the game	Sports & Entertainment	Sports & Entertainment
Headlines whose final classification was changed due to participants’ ratings		
First on CNN: Biden administration shipping 44,000 pounds of Nestlé formula Thursday | CNN Politics	Politics	US & Local News
Virginia parents call out ’political agenda’ as school board approves suspensions for ’misgendering’	Politics	US & Local News
Biden administration eases terrorism-related restrictions for Afghan evacuees	Politics	World News
Inside the South African company making some of America’s rarest and most beloved cars	World News	Economy & Business
Exercise pill? Researchers identify molecule in blood produced during workout	Health & COVID-19	Science & Technology

Table 4.4. Examples of headlines from each topic, and headlines whose final topic classification was changed.

Attention Checks on Mechanical Turk

Of the 276 participants that failed at least one attention checks, 120 did not give a correct response to the initial screener question "Puppy is to dog as kitten is to?", and were therefore redirected to the end of the study automatically without being given any news headlines. 35 other participants failed a second screener after responding to all 10 headlines; this question asks "What is your favorite color?", but instructs participants to always choose red and green. These responses were filtered out manually.

The remaining 121 participants did not provide a proper response to an open-ended

Headline	Actual veracity	Veracity score
True headlines that participants deemed as true		
Democratic lawmakers call for unisex bathrooms at the U.S. Capitol	True	5.81
McCaul: US must ’wake up’ and invest in Latin America to gain competitive edge on China	True	5.57
Tesla raises model prices amid global supply-chain shortage	True	5.47
FDA advisers vote in favor of authorizing COVID-19 vaccines for children as young as 6 months	True	5.37
True headlines that participants deemed as false		
Tom Brady plans to produce more movies after retirement: ’I definitely see that as part of my future’	True	3.07
Lauren Boebert taking legal action over Dem PAC’s ’false and disgusting claims’ that she was ’paid escort’	True	3.23
COVID vaccines for infants arrive in Florida. Here’s why doctors are throwing them away	True	3.31
Alaska kids served floor sealant instead of milk at elementary school summer program	True	3.34
Misinformation headlines that participants deemed as true		
Agents denying Trump freakout claim were his ’yes men’: WaPo	Misinformation	5.13
Why did Texas ban certain Instagram filters?	Misinformation	4.83
A news outlet asked every Democrat Senator if they’ll endorse Biden in 2024 - not many said yes	Misinformation	4.75
Texas bill would allow a rapist to sue his victim for having an abortion	Misinformation	4.73
Misinformation headlines that participants deemed as false		
John Stockton claims to have list of hundreds of vaccinated athletes who have dropped dead on the field	Misinformation	2.71
President Obama says, "Gotta have them ribs and p*ssy too!"	Misinformation	2.76
Official study concludes that masks caused more COVID	Misinformation	3.18
Military arrests Biden’s Sec. of Agriculture Tom Vilsack	Misinformation	3.20

Table 4.5. Examples of true headlines and misinformation headlines with high and low crowd-sourced veracity scores. Average veracity score is 4.45.

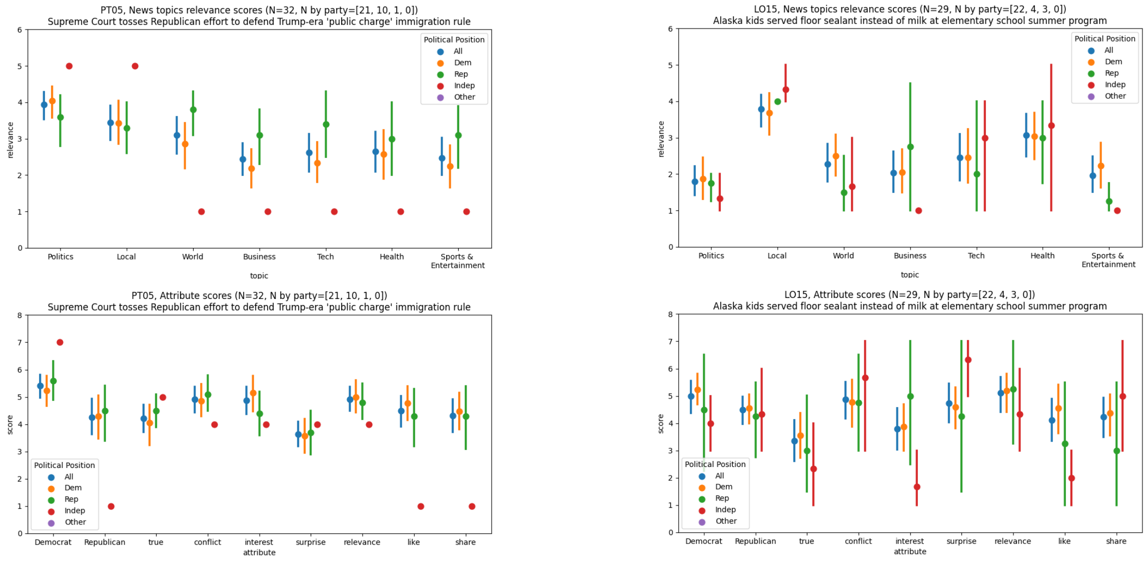
question. There are two such questions in the survey: an attention check after the fifth headline that requires participants to recall one of the headlines they have read and enter at

Headline	Score
Most and least favorable to Democrats	
Kamala Harris launches new national task force on preventing online harassment and abuse	6.06
Biden shakes hands with thin air after North Carolina speech	3.69
Most and least favorable to Republicans	
A STOLEN ELECTION: State Totals Minus Illegal Ballot Trafficking Numbers Give President Trump Decisive Victories in AZ, GA, MI, PA, and WI	5.71
Texas bill would allow a rapist to sue his victim for having an abortion	3.41
Most and least likely to be true	
Democratic lawmakers call for unisex bathrooms at the U.S. Capitol	5.81
John Stockton claims to have list of hundreds of vaccinated athletes who have dropped dead on the field	2.71
Most and least controversial or conflicting	
Exposed !! Pfizer CEO says it’s their dream to reduce the population by 50 percent in 2023 !!	5.91
When did Tiger Woods have the yips and how did he overcome it?	2.95
Most and least interesting or funny	
Delta pilots say they’ve been flying ’record amount of overtime’ amid flight cancellations	5.44
Tesla raises model prices amid global supply-chain shortage	2.80
Most and least surprising or unexpected	
U.S. to propose rule to limit nicotine levels in cigarettes	5.36
GOP Gov. Hutchinson says Trump responsible for Jan. 6 ’politically, morally’ but not criminally	2.79
Most and least relevant	
Democratic lawmakers call for unisex bathrooms at the U.S. Capitol	5.90
When did Tiger Woods have the yips and how did he overcome it?	3.10
Most and least likely to be liked	
Water-borne infections can lurk in hot tubs, public pools, lakes and oceans this summer: Here’s what to know	5.52
Bayer Executive Says mRNA Vaccines are Gene Therapy	3.00
Most and least likely to be shared	
The Marvel Cinematic Universe timeline on Disney+	5.53
Biden mumbles as he signs Executive Order ’Advancing LGBTQI+ Equality’ at reception for Pride Month (VIDEO)	2.80

Table 4.6. Examples of headlines with the highest and lowest scores on each attribute

least three keywords, and a final question at the end asking how much time the participant took to complete the survey.[13] Table 4.7 shows some responses from participants who have passed and failed the first question on recalling headlines.

[13] We did not keep track of the exact number of participants that failed each question, as there were overlaps.



(a) Supreme Court tosses Republican effort to defend Trump-era 'public charge' immigration rule

(b) Alaska kids served floor sealant instead of milk at elementary school summer program

Figure 4-5. Distributions of selected headlines' user ratings on topics (top) and attributes (bottom). Dots are averages, and vertical lines are 95% confidence intervals. Blue are distributions of responses from all users, and other colors are from users grouped by political leaning.

Accepted responses	Rejected responses
fuel price increase	political, economy, science
investigaiors, delta pilot, pharmacy	its a good work
Suspect, Accidentally, Stolen	Make the Headline Unique.
Moderna created the COVID pandemic.	one of the headlines you have read so far
Abbot baby formula plant agian stop production	Flush Left Headline. This is one of the more modern headline forms in use
CRYPTO COMPANIES, WATER BORNE INFECTIONS, COVID VACCINES	Political leaning,Veracity.Depictionof cinflict.
COVID 19 vaccines for kids under 5, Bill clinton, shopping stores is back and thriving	They can show how much progress is being made in economic terms

Table 4.7. Examples of responses to the headline recall attention check question

Wave 2 Hypotheses

While we do not have any results on Waves 1 and 2 yet, here are some hypotheses on what we might observe from user behavior towards an algorithmically ranked news feed in Wave 2:

- Friendships based on political ideology can intensify recommendation of misinformation and user engagement with them, but friendships based on other dimensions might dampen it (relative to a control group).

- Algorithmic ranking plays an important role in engagement with catchy click-bait content and misinformation. If this content is de-prioritized on the feed, even if it still exists but just falls lower on the feed, it will not be actively sought out.
- When algorithms recommend content shared by friends that also align with the user's own preferences, they result in greater impacts in user engagement with such recommended content, compared to the effects of recommendation by either factor alone.
- User engagement is lower beyond the "transition point" from tailored content to random content, once they notice the algorithm has run out of recommended articles, as compared to a fully randomized news feed (with no ranking algorithms) at the same location.

Chapter 5

Online Media and Taxing Digital Advertising

We present a model of digital advertising where users on a media platform (e.g., YouTube) consume both entertainment and advertisements (ads), which provide information about their preferences for a product. The platform chooses a business model that could allow a firm to advertise its product, and consumers make strategic choices about how much time to spend on the platform and how much of the product to buy. Our first main result shows that consumer welfare unambiguously decreases when the platform is monetized by digital ads, despite the information consumers gain from watching ads. Using this as motivation, we consider the impact of anti-trust regulation, with firm-level and platform-level competition as a potential corrective measure. Our second main set of results proves that competition can further decrease consumer welfare, because it might intensify platform incentives to target ads at susceptible populations who are most influenced. We conclude by recommending a solution addressing the heart of the problem, a digital advertising tax, which, if implemented well, will encourage platforms to switch from ad-based business models to subscription-based ones.

5.1 Introduction

> "Senior members from one of [Chile's] major political parties attributed their recent electoral success to their use of Facebook's targeted political ads . . . [Facebook] is a

> digital mercenary that is always the one with its finger on the trigger."
>
> — Paul Romer (2021)

In recent years, online platforms have become a dominant forum for entertainment and social interaction. The average adult spends over three hours a day on social media,[1] with significantly more time spent online through streaming services such as Netflix, YouTube, and Hulu (Budzinski et al. (2021), Richter (2019), and Twenge et al. (2019)). Possibly accelerated by the COVID-19 pandemic, there has been a steady transition from traditional media consumption (e.g., TV, print, or in-person interaction) to online media consumption over the last decade (see Cinelli et al. (2020) and Sherman and Waterman (2016)). While this transition has come with clear benefits, there is an ongoing debate about the potentially negative consequences lurking, including concerns about belief manipulation, mental health degradation, and digital addiction (respectively, see Marwick and Lewis (2017), Allcott et al. (2020), and Allcott et al. (2022)).

Business models of online platforms drive much of the content creation and algorithmic choices of platforms, and ultimately impact human-machine interactions. While some platforms generate revenue through other sources, a common business model for online media is digital advertising.[2] Unlike standard advertising, where the same product recommendation is broadcast to a large audience, digital advertising allows ads to be tailored and specifically targeted in ways that might make them more effective. There is a breadth of empirical literature documenting the existence and potential impacts of digital ad targeting (Bennett and Gordon (2020), Deng and Mela (2018), and De Jans et al. (2019)). Yet, we are currently devoid of a framework for assessing the welfare implications of digital ads and media platform business models in general. Without this framework, it is difficult to understand whether digital advertising poses a problem, and if so, what appropriate regulatory solutions could correct it.

In this paper, we develop a parsimonious model of an online media platform which can be monetized through advertising, subscriptions, or both. There is a firm with a horizontally-

[1] See `https://www.forbes.com/sites/petersuciu/2021/06/24/americans-spent-more-than-1300-hours-on-social-me` and `https://whatagraph.com/blog/articles/how-much-time-do-people-spend-on-social-media`.

[2] For example, digital advertising made up 98% of Facebook's revenue from 2017-2019 (see `https://www.nasdaq.com/articles/what-facebooks-revenue-breakdown-2019-03-28-0`) and about 85% of YouTube's revenue in 2020 (even with a premium ad-free subscription plan offered, see `https://spendmenot.com/blog/youtube-revenue-statistics/`).

differentiated product (some consumers may enjoy it, others may not) and a continuum of consumers who are (possibly) interested in purchasing the product. A media platform sits in between the firm and the consumers and provides dual services to its potential users. First, it offers entertainment through the content it supplies, which provides a known value to the consumer. Second, it can inject advertisements for the firm's product, which provide information to the consumer about whether she should purchase it. The consumer does not get entertainment value from the ad, but internalizes the informational gain from making a better-informed purchasing choice.

Ads are targeted at users based on a platform algorithm (e.g., calibrated to past online behavior such as video clicks and engagement). Each ad gives off either a positive or negative signal about the product, but users get idiosyncratic ads tailored to them which may result in different impressions for different users. While a good signal ad indicates higher likelihood of good quality and a bad signal ad indicates higher likelihood of bad quality, both type-I and type-II errors are possible.

There are two types of consumers on the platform. The first type of consumer is *sophisticated*, and is aware that specific ads are being shown to convince her to purchase the product. She has a perfectly specified prior about the actual likelihood of good and bad signals from the ads conditional on her true (but *ex ante* unknown) preference for the product. The second type of consumer is *naive* and unaware of any digital ad targeting. She has a misspecified prior about the distribution of signals from ads, and generally believes good signals are stronger than they actually are. This makes advertising particularly effective on naive agents, as it tends to be more convincing of the firm's product.

The platform can offer two services to participate in: (1) a free advertising-based plan that will have occasional ads for the firm's product, and (2) an ad-free plan that will directly charge the consumer a subscription fee. We first assume that the platform cannot implement both simultaneously, and must choose between the two. Our first main result shows consumer welfare strictly decreases when the platform monetizes according to advertising, because even though consumers learn about the product (and their preferences), the firm uses the advertising to raise prices, while the platform spams them with excessively many ads. Consumer welfare falls even further when the population of sophisticates is too low, because the platform will further extort naives by increasing the ad intensity without losing their participation. In this way, sophisticated agents in the population serve as a positive externality to protect society and help mitigate the negative welfare implications of digital ads.

Next, we allow the platform to adopt freemium-type pricing models, where both free advertising-based and ad-free subscription-based plans are available. In this case, consumer welfare generally degrades even further because a natural separating equilibrium emerges — digital ads can be used to prey specifically on naives while allowing sophisticates (who are not heavily influenced anyway) to opt into a subscription-based plan that is user-monetized. This consumer welfare degradation is increasing in how easily naives are duped by their targeted ads, but is non-monotone in the overall informativeness of ads. The latter is observed because ad informativeness may be more or less advantageous to users who can use ads to make better decisions, or to firms who can use this technology to extract more consumer surplus through advertising.

These observations encourage us to then turn to potential remedies for combating the negative welfare impacts of digital advertising. A commonly discussed approach is antitrust regulation that would break up the firms who advertise on media platforms (and set monopolistic prices for their products) or to break up the platforms themselves (who have total control over their means of revenue). We find that under the conditions where consumer welfare is already low (and the population consists largely of naives), firm-level and competition-level competition are not just ineffective at correcting the problem, they often exacerbate it. In the case of firm-level competition, advertising is used to secure a loyal user base and can result in higher prices compared to monopolistic levels. In turn, this further encourages platforms to adopt business models that focus on advertising-based revenue streams with an emphasis on targeting naive users who are more influenced by ads. On the other hand, platform-level competition hurts naives because subscription price competition makes freemium pricing a more stable outcome. When multiple platforms compete, there is a temptation to "steal" users who would prefer subscription, which leads to much more lenient conditions under which the freemium model (with the lowest welfare) is supported in equilibrium.

We conclude the paper with what we believe to be the most promising solution, a digital advertising tax. Such a proposal involves taxing platforms directly on the revenue earned from digital advertising sources, with the goal of pushing them toward other business models such as subscription. Our final result shows that such a tax can be effective if strongly implemented, but that light digital ad taxes can have either no effect or even decrease welfare further. In particular, light ad tax policies might have the same unintended negative consequences highlighted before, where it encouraged platforms to switch to a freemium pricing model with the goal of more aggressively targeting naives. However, stronger ad tax policies (such as a progressive ad tax) can fully correct platform incentives and push to fully restoring maximal consumer welfare.

Related Literature. Our paper builds on a much broader work centered around advertising, online platforms, and consumer welfare. Along with the literature mentioned previously, several other papers are related to our findings.

There is a rich literature dating prior to the digital era that studies traditional advertising (e.g., see Tirole (1988) and Grossman and Shapiro (1984)). Most closely related to our digital advertising setup is the work of Meurer and Stahl (1994), building on the seminal paper of Butters (1977). In their model, advertising is informative about partially substitutable, horizontally-differentiated products, and consumers use the information in ads to decide which products to buy. We differ on three important dimensions. First, their model has no platform, and consumers therefore make no active decisions about whether to engage in ad viewership or not. Second, digital ads (as opposed to traditional ads) have the potential to target naive populations, which we model as different from the standard sophisticated agent benchmark. Our key results turn on this property of the model. Lastly, at the core of our findings is how the platform's decision of business model affects consumer welfare, which is absent in Meurer and Stahl (1994).

More recent work looks at data sharing and the implications for targeted digital advertising. Acemoglu et al. (2022a) identify a market inefficiency in how users price their data due to data sharing and leaks between platforms. This leads to lower consumer welfare, and as in our findings, increasing competition among platforms may fail to remedy the issue. The paper of Marotta et al. (2021) takes a similar approach in the context of digital advertising, showing that data sharing can lead to more effective targeting of ads at consumers, resulting in potentially worse outcomes. In our paper, we abstract away from the precise targeting technology by adopting naive agents who are overly influenced by tailored ads. This simplification allows us to effectively study the optimal business model of the platform and its impacts on welfare, leading to new insights and policies that can redress certain negative platform incentives.

While online media platforms are fairly new, there is also a recent strand of literature that studies platform algorithmic choices and monetization. Sato (2019) find that the optimal business model of a digital platform is at most a two-item menu, often known as “freemium”, with a free ad-based plan and a paid-for premium plan with no ads. There are also various empirical studies demonstrating the profitability and success of freemium business models in settings such as online streaming and social media gaming (see Montag et al. (2019), Rietveld (2018), and Holm and Günzel-Jensen (2017)). A central focus of this paper is to study how platforms choose their “freemium” menu and the implications for consumer welfare.

There are also papers that study how platform incentives and algorithms might otherwise negatively impact user well-being, such as through promoting misinformation (Acemoglu et al. (2022b)) or by heightening personal insecurities (Allcott et al. (2020)). These works serve as complementary to this work, which provides a framework for studying the effects of platform business models and monetization methods on consumer well-being.

The rest of the paper is organized as follows. The next section introduces our modeling environment, describes agent payoffs, and defines consumer welfare. Section 5.3 characterizes the unique (Berk-Nash) equilibrium of the model and provides some comparative statics of interest. Section 5.4 generalizes the baseline model to allow the platform to adopt richer (mixed) business models, with subtler implications for consumer welfare. Section 5.5 studies the effects of introducing firm-level and/or platform-level competition. Section 5.6 concludes by discussing the proposed solution of a digital ad tax to correct platform incentives that lead to worse business models for consumers.

5.2 A Model of Content Platforms

Our baseline model consists of three types of agents: a single firm, a single media platform, and a continuum of consumers. The firm is a monopolist who sells a single product. The media platform supplies entertainment to its users, but can intermix advertisements (from now on, simply ads) that market the product by providing information to the consumer about her preferences (for that product). The consumers are both potential users of the platform and potential purchasers of the product. In our baseline model, we allow the platform to choose whether to be monetized entirely by ad revenue or entirely by subscription fees.[3]

[3] In Section 5.4, we generalize this model to allow the platform to offer both types of plans (ad-based or subscription-based). This business model commonly arises in streaming services (e.g., YouTube or Hulu) where

Consumers. There is a continuum of consumers who each have a two-dimensional type $(\tau_i, \theta_i) \in \{S, N\} \times \{0, 1\}$. The first dimension of the consumer's type corresponds to her sophistication; each consumer i is either sophisticated ($\tau_i = S$) or naive ($\tau_i = N$). Sophisticated consumers are immune to digital ad targeting but naive consumers are susceptible to them. Each consumer is sophisticated with probability λ and naive with probability $1-\lambda$ (independent across consumers). The second dimension of the consumer's type aligns with an unknown attribute θ_i of whether she would get positive utility from using the product; each consumer i either likes the product ($\theta_i = 1$) or does not like the product ($\theta_i = 0$), although this attribute is not known to consumer i *ex ante*. The consumer has ex-ante probability q that she will like the product (independent across consumers).

Each user gets a series of personalized advertisements that provides her with information about her true preferences. Formally, each ad to user i provides a binary signal $s_i \in \{\ominus, \oplus\}$ about the product, which is independent across ads and consumers. The signal distribution for an ad is given by:

$$\begin{cases} s_i = \oplus, & \text{with probability } \phi_1 \text{ if } \theta_i = 1 \\ s_i = \ominus, & \text{with probability } 1 - \phi_1 \text{ if } \theta_i = 1 \\ s_i = \oplus, & \text{with probability } \phi_0 \text{ if } \theta_i = 0 \\ s_i = \ominus, & \text{with probability } 1 - \phi_0 \text{ if } \theta_i = 0 \end{cases} \tag{5.1}$$

where we assume that $\phi_1 > \phi_0$. Therefore, a positive signal $s_i = \oplus$ provides information to the consumer that she is *more* likely to like the product, whereas a negative signal $i = \ominus$ provides information to the consumer that she is *less* likely to like the product. This signal distribution of Equation (5.1) the *objective* model, where both type-I and type-II errors are possible; that is, an agent i with $\theta_i = 1$ might receive a negative signal $s = \ominus$ or an agent i with $\theta_i = 0$ might receive a positive signal $s = \oplus$. The former could occur if a particular ad is off-putting to a certain demographic (e.g., an unfunny insurance commercial), whereas the latter might happen if the ad uses tangential but attractive images to glamorize the product (e.g., soft drinks at the beach). However, the assumption that $\phi_1 > \phi_0$ implies that ads are at least partially informative about preference.

there is an ad-free experience at a premium charge to the consumer.

Sophisticated agents have a perfectly specified model but naive agents have a misspecified model. In the naives *subjective* model, advertising signals are believed to be generated according to

$$\begin{cases} s = \oplus, & \text{with probability } \phi_1 \text{ if } \theta_i = 1 \\ s = \ominus, & \text{with probability } 1 - \phi_1 \text{ if } \theta_i = 1 \\ s = \ominus, & \text{always if } \theta_i = 0 \end{cases}$$

instead of the objective model of Equation (5.1). In other words, a naive agent i believes that $s = \ominus$ for every ad whenever $\theta_i = 0$, so any positive signal $s = \oplus$ suggests to her that she would indeed like the product. In other words, a naive agent is overly optimistic about any ad that appeals to her.

We use the notion of Berk-Nash equilibrium (Esponda and Pouzo (2016)) to model agents' beliefs with misspecified priors. Observe that because the subjective model of sophisticates *is* the objective model, a sophisticated agent will have a standard Bayesian belief π^S about θ_i conditional on seeing k_+ positive ads and k_- negative ads:

$$\pi^S = \frac{\phi_1^{k_+}(1-\phi_1)^{k_-}q}{\phi_1^{k_+}(1-\phi_1)^{k_-}q + \phi_0^{k_+}(1-\phi_0)^{k_-}(1-q)}$$

Because the subjective model of naives *differs* from the objective model, a naive agent's belief π_i will be the one that minimizes divergence between her observation of ads (k_+, k_-) and her subjective model. Note, however, that every realization of (k_+, k_-) is *consistent* with her subjective model and the divergence can be minimized at 0 with the following update:

$$\pi^N = \frac{\phi_1^{k_+}(1-\phi_1)^{k_-}q}{\phi_1^{k_+}(1-\phi_1)^{k_-}q + \mathbf{1}_{k_+=0}(1-q)}$$

This pins down the beliefs of both types of agents conditional on observing k_+ positive ads and k_- negative ads.

Firm. There is a monopolist who sells a single product. The firm is allowed to set a unit price p for the product, whereby any consumer i who purchases z_i pays price pz_i. The firm can produce their product at a constant marginal cost c.

Platform. The platform shows content (e.g., videos) to the users but can determine the ratio of ads to total content (entertainment plus ads), α. Every ad seen by a consumer provides some information about consumer i's type θ_i, as described above.

5.2.1 Actions and Timing

Next, we define a strategic game with all three types of agents. The game will consist of five time periods, denoted $t = 1, 2, 3, 4, 5$, as depicted in Figure 5-1.

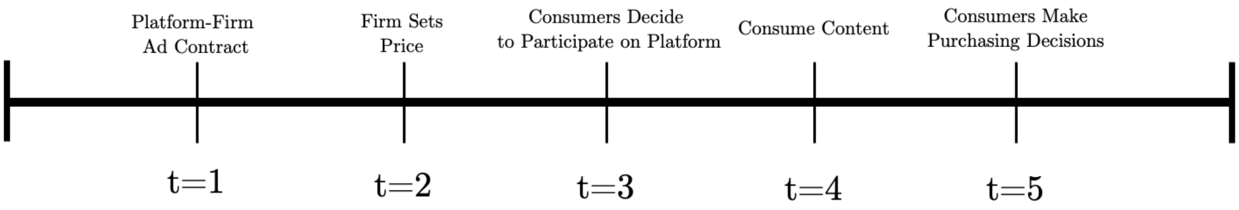

Platform-Firm Ad Contract
Firm Sets Price
Consumers Decide to Participate on Platform
Consume Content
Consumers Make Purchasing Decisions
t=1
t=2
t=3
t=4
t=5

Figure 5-1. Timing of the Advertising Model.

(i) At $t = 1$, the platform and the firm negotiate a contract that specifies an ad intensity α and a monetary transfer m from the firm to the platform. For simplicity, we assume this takes the form of a take-it-or-leave-it offer (α, m) from the platform to the firm (that the firm either accepts or rejects). If the firm rejects the contract (or the platform does not offer one), the platform can set a subscription fee P to charge users who want to participate directly.

(ii) At $t = 2$, the firm sets its price p^* for its product.

(iii) At $t = 3$, the platform produces content and advertises at rate α. If the contract is accepted, each consumer i makes a binary decision $x_i \in \{0, 1\}$ about whether to allocate time T ($x_i = 1$) or no time ($x_i = 0$) on the platform, given α.[4] If the contract is rejected, each consumer i makes the same binary decision $x_i \in \{0, 1\}$, given the subscription fee P. A decision $x_i = 0$ to not participate gives user i her outside option v.

(iv) At $t = 4$, the consumer digests the platform content (including any ads). Any consumer who engages with the platform ($x_i = 1$) views $\lceil \alpha T \rceil$ ads and receives $T - \lceil \alpha T \rceil$ time of

[4] For simplicity and ease of exposition, we assume the consumer decides whether to allocate some (exogenously given) time T or allocate no time at all to consuming platform content. We note, however, that our results are not sensitive to a continuous time allocation decision $x_i \in [0, 1]$ for each consumer.

entertainment. Any consumer who does not engage with the platform ($x_i = 0$) views no content whatsoever (ads or entertainment).

(v) At $t = 5$, each consumer i decides how much of the product to purchase, z_i, at the price p.

5.2.2 Payoffs and Solution Concept

Next, we describe the payoffs that each type of agent receives.

Platform. In our baseline model, the platform can generate revenue by charging the firm for advertising *or* by charging users a subscription fee. The platform's payoff is thus m^* if a contract (α^*, m^*) is accepted by the firm and otherwise is $\int_0^1 P^* x_i \, di$ where P^* is the subscription fee set by the platform.

Firm. The firm generates a profit by selling its product, but pays the platform for advertising. That is, the firm receives a payoff $\int_0^1 (p^* - c) z_i^* \, di - m^*$, where z_i^* is the consumption decision of agent i and m^* is the agreed upon transfer for an accepted contract (and zero for a rejected contract).

Consumers. Each consumer i receives utility both from product consumption and from content consumption on the platform. A consumer i with type θ_i receives a consumption utility $U(z_i; \theta_i) = \beta \theta_i z_i - z_i^2/2$ for purchasing z_i units of the product, with total payoff given by $U(z_i; \theta_i) - p z_i$. Here, the parameter β represents the relative size of the product market compared to the market for entertainment.[5] Each consumer receives a direct utility $T - \lceil \alpha T \rceil$ from content enjoyment on the platform (if $x_i = 1$) and otherwise receives an outside option $v \geq 0$ from alternative activities (outside the platform).[6] We assume $v < T$, so an ad-free platform provides higher utility to the consumer than her outside option.

Solution Concept. Our solution concept is Berk-Nash equilibrium (see Esponda and Pouzo (2016)) of the aforementioned sequential game, which can be determined via backward

[5] As we discuss in Section 5.3, β roughly captures the ratio of the "willingness-to-pay" of advertisers to the "willingness-to-pay" of content consumers. Empirical measures of these on various digital platforms have been studied recently in works such as Berger et al. (2015), Sherman and Waterman (2016), and Flew (2021).

[6] The outside option v can also be used to model content platforms where the ad-based experience is still not free (e.g., with the streaming service Hulu). If v' is the true outside option and f is the fee charged for an ad-based plan, then the model can be applied identically by using outside option $v \equiv v' + f$.

induction.[7]

(a) At $t = 5$, each consumer holds belief π_i that she values the product ($\theta_i = 1$) and selects her optimal consumption z_i^* to solve $C_i(\pi_i, p) \equiv \max_{z_i} \pi_i U(z_i; \theta_i = 1) + (1 - \pi_i)U(z_i; \theta_i = 0) - pz_i$. Here, $C_i(\pi_i, p)$ represents the expected consumption utility ***given*** a belief π_i about $\theta_i = 1$ and a given product price p.

(b) At $t = 4$, given consumer i's choice of participation x_i and the ad intensity α, the consumer forms belief π_i (determined by her sophistication type τ_i and unknown preference type θ_i).

For $t \leq 3$, there are two separate subgames. In the **advertising subgame**, when an advertising contract is accepted:

(c) At $t = 3$, each consumer solves $\max_{x_i} x_i(\mathbb{E}_{\pi_i}[C_i(\pi_i, p) \mid \alpha] + T - \lceil \alpha T \rceil) + (1 - x_i)(C_i(q, p) + v)$. That is, the consumer chooses $x_i \in \{0, 1\}$ (whether to participate or not) based on the expected consumption utility C_i and her direct utility from using the platform, conditional on the ad intensity α. Let $x_i^*(\alpha, p)$ denote the platform participation decision of consumer i, as a function of the ad intensity α and product price p.

(d) At $t = 2$, given ad intensity α, the firm sets price p by solving $\Pi(\alpha) \equiv \max_p \int_0^1 x_i^*(\alpha, p)(p - c)\mathbb{E}_{\pi_i}[z_i^*(\pi_i, p) \mid \alpha] + (1 - x_i^*(\alpha, p))(p - c)z_i^*(q, p)\, di$. Here, $\Pi(\alpha)$ is the profit the firm receives from sales of the product, given an advertising intensity α on the platform.

(e) At $t = 1$, given a contract (α, m), the firm accepts if and only if $\Pi(\alpha) - m \geq \max_p (p - c)z_i^*(q, p)$ (i.e., the contract yields more profit than under the benchmark of zero advertising). The platform then selects the contract (α, m) that maximizes m conditional on acceptance by the firm.

In the **subscription subgame**:

(c) At $t = 3$, each consumer solves $\max_{x_i} x_i(T - P) + (1 - x_i)v$.[8] Let $x_i^*(P)$ denote the platform participation decision of consumer i, as a function of the subscription fee P.

[7] As is standard, the equilibrium can be solved by starting at $t = 5$, taking all quantities determined at $t \leq 4$ as given, and solving for the agents' best-response actions. From there, one can solve for the best responses in $t = 4$ given the best-response correspondence already pinned down at $t = 5$ and quantities taken as given in $t \leq 3$. Repeating this until $t = 1$ gives the unique Berk-Nash equilibrium.

[8] Observe that consumption utility does not factor into the consumer's decision in a subscription-based platform. Because the platform does not provide any information about the product in the form of advertising, the platform participation decision and the consumption decision are completely uncoupled.

(d) At $t = 2$, the firm sets price p by solving $\max_p (p - c) z_i^*(q, p)$.

(e) At $t = 1$, the platform chooses a subscription service with fee P to maximize $\int_0^1 P x_i^*(P)\ di$.

Finally to determine, which subgame is taken on the equilibrium path, the platform compares the profits from (e) in the advertising subgame and from (e) in the subscription subgame and picks the more profitable business model.

5.2.3 Consumer Welfare

We first start with the base case of consumer welfare when no platform exists at all.[9] It is straightforward to see here that optimal consumption satisfies $z_i^*(q, p^*) = \beta q - p^*$ and the firm will set a price of $p^* = (\beta q + c)/2$ in equilibrium. Average consumer welfare is then given by $U(z_i^*(q, p^*); \theta_i = q) - p^* z_i^*(q, p^*) + v = (\beta q - c)^2/8 + v$. This welfare is the same for both sophisticated and naive agents because in the base case, there is no advertising, and both types of agents take identical actions with identical payoffs.

When there is a media platform that may advertise, we can measure consumer welfare as follows. First, suppose that (α^*, m^*) is the contract accepted and p^* is the price chosen by the firm in equilibrium. We let F_{S0}, F_{S1}, F_{N0} and F_{N1} denote the distributions over π_i for the respective types, $(S, 0)$, $(S, 1)$, $(N, 0)$, and $(N, 1)$.[10] Note that for π^* and p^* in equilibrium at $t = 5$, there is a unique optimal consumption choice $z^*(\pi^*, p^*)$ independent of type. Thus, average

[9] We remark that consumer welfare under a different base case, where the platform exists but *must* be monetized through subscriptions, is identical in our model. In this setting, a lack of advertising means the platform provides no information about the firm's product, so their welfare in the product market is the same as without a platform. At the same, the platform will charge a subscription fee that extracts all "entertainment" surplus from the consumer, so the platform provides no added utility to the consumer, and overall consumer welfare remains unaffected.

[10] Formally, these are given by the multinomial distribution:

$$\mathbb{P}_{S\theta_i}\left[\pi_i = \frac{\phi_1^{k_+}(1-\phi_1)^{k_-}q}{\phi_1^{k_+}(1-\phi_1)^{k_-}q + \phi_0^{k_+}(1-\phi_0)^{k_-}(1-q)}\right] = \binom{k_+ + k_-}{k}\phi_{\theta_i}^{k_+}(1-\phi_{\theta_i})^{k_-}$$

$$\mathbb{P}_{S\theta_i}\left[\pi_i = \frac{\phi_1^{k_+}(1-\phi_1)^{k_-}q}{\phi_1^{k_+}(1-\phi_1)^{k_-}q + \mathbf{1}_{k_+=0}(1-q)}\right] = \binom{k_+ + k_-}{k}\phi_{\theta_i}^{k_+}(1-\phi_{\theta_i})^{k_-}$$

for all $0 \leq k \leq \lceil \alpha T \rceil$, and 0 for all $k > \lceil \alpha T \rceil$. Here, recall that k_+ are the number of positive signals and $k_- = \lceil \alpha T \rceil - k_+$ negative signals.

welfare by type (conditional on ad-based platform engagement) is given directly by

$$\begin{aligned} W(\tau_i, x_i = 1) = \underbrace{T - \lceil \alpha^* T \rceil}_{\text{Content Consumption Utility}} &+ q \underbrace{\mathbb{E}_{\pi \sim F_{\tau_i 1}} \left[U(z^*(\pi, p^*);\, \theta_i = 1) - p^* z^*(\pi, p^*) \right]}_{\text{Product Consumption Utility Conditional on } \theta_i = 1} \\ &+ (1-q) \underbrace{\mathbb{E}_{\pi \sim F_{\tau_i 0}} \left[U(z^*(\pi, p^*);\, \theta_i = 0) - p^* z^*(\pi, p^*) \right]}_{\text{Product Consumption Utility Conditional on } \theta_i = 0} . \end{aligned} \tag{5.2}$$

Second, suppose that the platform offers a subscription price P^* instead. Then, average welfare by type (conditional on platform engagement) is

$$W(\tau_i, x_i = 1) = \underbrace{T - P^*}_{\text{Content Consumption Utility}} + \underbrace{qU(z_i^*(q, p^*);\, \theta_i = 1) + (1-q) U(z_i^*(q, p^*);\, \theta_i = 0) - p^* z_i^*(q, p^*)}_{\text{Expected Product Consumption Utility}} .$$

And finally, when the user abstains from participating on the platform, the welfare is given simply by

$$W(\tau_i, x_i = 0) = \underbrace{v}_{\text{Outside Option}} + \underbrace{qU(z_i^*(q, p^*);\, \theta_i = 1) + (1-q) U(z_i^*(q, p^*);\, \theta_i = 0) - p^* z_i^*(q, p^*)}_{\text{Expected Product Consumption Utility}} . \tag{5.3}$$

Thus, in equilibrium, we can define welfare $W^*(S) = W(S, x_S^*)$ and $W^*(N) = W(N, x_N^*)$ for each agent sophistication type (noting of course the platform engagement in equilibrium will be the same for all agents of the same sophistication type). Finally, we can also define *average consumer welfare* of the entire society as $\bar{W}^* = \lambda W^*(S) + (1-\lambda) W^*(N)$.

5.3 Equilibrium Characterization

We next characterize the unique equilibrium of our advertising game and the consumer welfare from product and content consumption with digital ads. At $t = 5$, one can solve directly to see that user i's optimal consumption is given by $z_i^*(\pi_i, p) = \beta\pi_i - p^*$ and yields expected utility $C_i(\pi_i, p^*) = (\beta\pi_i - p^*)^2/2$ for the consumer. At $t = 3$, we see that a consumer engages with an advertising platform if and only if

$$\underbrace{q\mathbb{E}_{\pi_i|\theta_i=1}[C_i(\pi_i,p^*)\mid\alpha^*]+(1-q)\mathbb{E}_{\pi_i|\theta_i=0}[C_i(\pi_i,p^*)\mid\alpha^*]-C_i(q,p^*)}_{\text{Information Gain}}+\underbrace{T-\lceil\alpha^*T\rceil-v}_{\text{Net Platform Enjoyment}}\geq 0\,.$$

One can observe the informational gain simplifies to $q\mathbb{E}_{\pi_i|\theta_i=1}[(\beta\pi_i)^2/2\mid\alpha^*]+(1-q)\mathbb{E}_{\pi_i|\theta_i=0}[(\beta\pi_i)^2/2\mid\alpha^*]-(\beta q)^2/2$, which depends on α^* and the sophistication type τ_i of agent i (as the subjective distribution of $\pi_i|\theta_i$ depends on τ_i), but not the price p^*. That is, the information gain from digital ads for any agent i depends only on the ad intensity and not the price charged by the firm. For ease of notation, we denote this informational gain by $I_\tau(\alpha^*)$ for any user of type $\tau\in\{S,N\}$. On the other hand, for a subscription platform, the expected consumption utility is independent of platform participation. Hence, participation is a best response if and only if $T-P^*-v\geq 0$ for both types of users.

The decision of the consumer to participate on an advertising platform depends on two factors. First, the consumer internalizes the information gain from using the platform, which comes in the form of advertisements that inform the consumer about her true type θ_i. Second, the user receives direct utility (in the form of entertainment) from consuming content, which is compared to her outside option. If the sum of these expressions is non-negative, the consumer finds platform participation to be a best response.

For $t=1$ and $t=2$, we can collapse the decision problem of the platform and firm because the platform offers a take-it-or-leave-it offer. In particular, the platform will solve a set of maximization problems with various participation constraints. Let us denote by $\bar{\pi}_F$ the expected belief under distribution F, i.e., $\bar{\pi}_F = \mathbb{E}_{\pi\sim F}[\pi]$. Then the platform solves

$$\mathcal{A}_1 \equiv \max_{\alpha,p} \;(p-c)(\beta q - p) \qquad \text{(No user participation)}$$

$$\begin{aligned} \mathcal{A}_2 \equiv \max_{\alpha,p} \;& \lambda(p-c)(\beta q \bar{\pi}_{F_{S1}(\alpha)} + \beta(1-q)\bar{\pi}_{F_{S0}(\alpha)} - p) + (1-\lambda)(p-c)(\beta q - p) \qquad \text{(Sophisticates participate} \\ \text{subject to} \;& I_S(\alpha) + T - \lceil \alpha T \rceil - v \geq 0 \end{aligned}$$

$$\begin{aligned} \mathcal{A}_3 \equiv \max_{\alpha,p} \;& (1-\lambda)(p-c)(\beta q \bar{\pi}_{F_{N1}(\alpha)} + \beta(1-q)\bar{\pi}_{F_{N0}(\alpha)} - p) + \lambda(p-c)(\beta q - p) \qquad \text{(Naives participate)} \\ \text{subject to} \;& I_N(\alpha) + T - \lceil \alpha T \rceil - v \geq 0 \end{aligned}$$

$$\begin{aligned} \mathcal{A}_4 \equiv \max_{\alpha,p} \;& (p-c)(\lambda\beta q \bar{\pi}_{F_{S1}(\alpha)} + \lambda\beta(1-q)\bar{\pi}_{F_{S0}(\alpha)} \\ & + (1-\lambda)\beta q \bar{\pi}_{F_{N1}(\alpha)} + (1-\lambda)\beta(1-q)\bar{\pi}_{F_{N0}(\alpha)} - p) \qquad \text{(All users participate)} \\ \text{subject to} \;& I_S(\alpha) + T - \lceil \alpha T \rceil - v \geq 0 \\ & I_N(\alpha) + T - \lceil \alpha T \rceil - v \geq 0 \end{aligned}$$

We can further simplify the platform's problem by noting that it is without loss to restrict attention to just $\mathcal{A}_3$ and $\mathcal{A}_4$. First, one can observe that $q\bar{\pi}_{F_{S1}(\alpha)} + (1-q)\bar{\pi}_{F_{S0}(\alpha)} = q$ because sophisticated agents have a properly specified Bayesian model, and thus, $\mathcal{A}_1 \geq \mathcal{A}_2$. At the same time $\bar{\pi}_{F_{N1}(\alpha)} > \bar{\pi}_{F_{S1}(\alpha)}$ and $\bar{\pi}_{F_{N0}(\alpha)} > \bar{\pi}_{F_{S0}(\alpha)}$, so $q\bar{\pi}_{F_{N1}(\alpha)} + (1-q)\bar{\pi}_{F_{N0}(\alpha)} > q$. We also note that $I_N(\alpha) + T - \lceil \alpha T \rceil - v \geq 0$ can be feasibily satisfied at $\alpha = 0$ (by assumption that $T > v$), so it must be that $\mathcal{A}_3 \geq \mathcal{A}_1 \geq \mathcal{A}_2$. That is, conditional on adopting an advertising model, the platform will advertise to attract both types of users or to attract only naives.

If $T - v > \max\{\mathcal{A}_3, \mathcal{A}_4\}$, the platform adopts a subscription model instead, with fee $P^* = T - v$ (extracting all consumer surplus). Otherwise, the platform adopts an advertising model, choosing the largest between $\mathcal{A}_3$, and $\mathcal{A}_4$, with the ad intensity α^* that maximizes that respective expression. In this contract, the platform chooses the transfer m that extracts all excess profits from the firm, which will set p^* as the price that solves the same joint optimization problem.[11]

[11] A more realistic model might be one where the firm and platform bargain over a contract and each takes some positive surplus ala Nash bargaining. Our assumption that the platform has all of the bargaining power leads to an identical equilibrium with the exception of the split m in the contract, which is immaterial to our consumer welfare analysis.

5.3.1 Optimal Product Pricing and Ad Intensity

To characterize the solution to the platform's problem, we leverage the following observation.

Lemma 5.3.1. *For any α, $I_N(\alpha) > I_S(\alpha) > 0$. Moreover, $I_S(\alpha)$ and $I_N(\alpha)$ are monotonically increasing in α.*

The intuition for Lemma 5.3.1 is as follows. Because naives mistakenly believe that the information contained in ads is more valuable than it is, their anticipated gain from participating and learning about θ_i is higher than that of sophisticates. Consequently, naives are more tolerant of ads because of the incorrect belief that it will improve their expected utility from product consumption more than it does. This implies that the platform participation constraint will always bind for sophisticates before it binds for naives, so whenever sophisticates participate, so do naives. Formally, it allows the platform to drop the naives' participation constraint in $\mathcal{A}_4$.

Secondly, one can show that the objectives of $\mathcal{A}_3$ and $\mathcal{A}_4$ are monotonically increasing in α. This observation and Lemma 5.3.1 allows the platform to first optimize over α given the separate binding participation constraint in each $\mathcal{A}_3$ and $\mathcal{A}_4$ (as a function of price p), then secondly optimize over price p^*. Finally, it compares the profits of $\mathcal{A}_3$, $\mathcal{A}_4$, and $T - v$. This solution is characterized in the following result.

Proposition 5.3.1. *There exists $\lambda^* \in (0, 1)$ and $\beta^* \in (0, 1)$ such that:*

(a) If $\beta < \beta^*$, the platform chooses a subscription model with $P^* = T - v$ and the firm sets a price $p^* = p_{sub}$;

(b) If $\lambda > \lambda^*$ and $\beta > \beta^*$, the platform chooses $\alpha_S^* = \sup\{\alpha \in [0, 1] : I_S(\alpha) + T - \lceil \alpha T \rceil - v \geq 0\}$ and the firm chooses a price $p_S^* > p_{sub}$;

(c) If $\lambda < \lambda^*$ and $\beta > \beta^*$, the platform chooses $\alpha_N^* = \sup\{\alpha \in [0, 1] : I_N(\alpha) + T - \lceil \alpha T \rceil - v \geq 0\} > \alpha_S^*$ and the firm chooses a price $p_N^* > p_S^*$.

There are three regimes identified in Proposition 5.3.1. The first is where the product market is not as sizable as the entertainment market ($\beta < \beta^*$). In this regime, the platform simply trades off what it can extract from firms versus what it can extract from content consumers, and rules in favor of simply charging users for its content. In practice, this does not seem to be the case for many media platforms (e.g., see Chyi (2005), Vock et al. (2013), and Chyi and

Ng (2020)). For example, for Facebook to match its advertising revenue via subscription, it would have to charge a subscription price around \$4/month. Recent evidence shows that most users would not be willing to pay this much, making the advertising model more lucrative for platforms.[12]

The other two regimes characterize the optimal advertising scheme given that the value of advertising to the firm outweighs the user's willingness-to-pay for entertainment. When a sizable fraction of the population is sophisticated, the platform does not want to alienate these consumers by imposing an ad intensity that is too high. Instead, the platform binds the participation constraint of the sophisticated agents, while naives (who are less bothered by ads) enjoy positive surplus from using the platform.

However, when sophisticates represent a smaller proportion of the population, the platform decides to only appeal to naives. In this case, sophisticates do not participate, and the platform increases its ad intensity to strip all utility of platform consumption from naives. At the same time, the firm charges a higher price for its product relative to regime (b), given that there is more advertising and the naive consumers are willing to pay higher prices after viewing more ads (on average). In this sense, in regime (b), the existence of sophisticates protects the naives from being extorted by excessive ads and high prices in the product market.

5.3.2 Welfare Analysis

We consider the welfare impacts from a platform monetized by ads or subscriptions relative to the base case where the platform does not exist at all, as described in Section 5.2.3. The existence of the platform provides the consumer with potential utility through two channels. First, it provides the consumer with information about her preference for the product (i.e., I_S or I_N) which can only be ascertained through viewing digital ads. Second, the platform provides entertainment, with supplies direct utility. However, the advertising allows the firm to generate higher profit by increasing its price and targeting those who desire the product more based on ad viewership.

Using the welfare definition of Section 5.2.3, our next result ranks the welfare of both sophisticated and naive users under Proposition 5.3.1 relative to the base case of no platform. As seen in Proposition 5.3.1, regimes (a), (b), and (c) all have different platform choices for α^*

[12] See https://omarzahran.medium.com/the-case-for-an-ad-free-social-media-subscription-c921eeeaaf7a.

and different firm choices for p^*, which is the unique equilibrium for the parameter values that satisfy the conditions of (a), (b), and (c), respectively. Here, we fix the underlyings of the model but compare welfare for individual agents under each of the (α^*, p^*) pairs of (a), (b), (c), and the base case.

We denote the base case consumer welfare as $W^*_{\text{base}}(\tau)$, which is the same for both types $\tau \in \{S, N\}$ and equal to $W^*_{(a)}(\tau)$ (see Footnote 9). In regime (b), we note that $x^*_i = 1$ for all users, and so we can use Equation (5.2) to compute welfare $W^*_{(b)}(S)$ and $W^*_{(b)}(N)$ for both sophisticates and naives, respectively. In regime (c), we have $x^*_i = 1$ for naives but $x^*_i = 0$ for sophisticates. Consequently, we can use Equation (5.2), as before, to compute welfare $W^*_{(c)}(N)$ for naives, but need to use Equation (5.3) to compute welfare $W^*_{(c)}(S)$ for sophisticates.

Theorem 5.3.1. *Consumer welfare for sophisticates and naives satisfies* $W^*_{base}(\tau) = W^*_{(a)}(\tau) > W^*_{(b)}(\tau) > W^*_{(c)}(\tau)$ *for both* $\tau \in \{S, N\}$.

Our first main result is a striking one. The introduction of a platform, which on the surface acts as a positive good to the consumers, providing both entertainment and information, necessarily reduces the consumer welfare of both types of agents. In essence, the additional value that the platform provides in digital ads (information) and entertaining content (enjoyment utility) are fully extracted by the firm and the platform, respectively, making them worse off than under the base case of no platform at all. Stated simply, platforms that use digital advertising as their main business model unambiguously hurt consumers.

Theorem 5.3.1 also emphasizes the importance of sophisticated consumers in protecting naive ones, as per Proposition 5.3.1(b). When a platform wants to broadly appeal to all types of consumers, it is forced to make digital ads less invasive and firms cannot as aggressively prey on those most influenced by them. As a result, welfare is better for all consumers when $\lambda > \lambda^*$, and the platform is forced to manage the aggression of its advertising.

Remark — The observed reduction in consumer welfare is not driven by the digital advertising technology itself, but by how the platform and firm use ads to boost profits at the expense of consumers. This is not to say that a social planner could not use the digital advertising technology in the same way to *benefit* consumers. For example, the social planner might be able to maximize consumer welfare (even above the base case) by introducing some ads (that provide information to consumers about the product) while simultaneously keeping product prices mostly unchanged.

5.4 Mixed Platform Business Models

Many media platforms are not monetized entirely from one revenue stream. Instead, the platform may allow the consumer to directly pay for an ad-free experience if she so desires (subscription) or participate for free but in the presence of ads (ad-based). This is often referred to as a *freemium* business model. In this section, we extend our baseline model of Section 5.2 to allow the platform to offer both subscription-based and ad-based plans that users can choose from.

5.4.1 Advertising, Subscription, and Freemium Models

We now suppose that in addition to the platform offering an ad contract (α, m) to the firm at $t = 1$, it can also announce a subscription plan at price P. Consumers at $t = 3$ now have an additional option conditional on $x_i = 1$: they can elect their plan $y_i \in \{\mathbb{A}, \mathbb{S}\}$, corresponding to whether to engage with the platform's content for free, but which contains ads (option $\mathbb{A}$), or to pay P to engage with the platform's content without ads via subscription (option $\mathbb{S}$), yielding payoff $T - P$ to the consumer and P to the platform.

This results in one of three considerations for the platform in equilibrium. First, it can choose a purely advertising-based model, as many social media sites do, with no option to avoid ads via subscription. If it does this, the platform chooses the ad intensity α_S^* as in Proposition 5.3.1(b).[13] Second, the platform can choose a purely subscription-based model, as some streaming services do such as Netflix (as of 2021), where consumers must pay directly to use the platform. Just as before, the platform will offer a subscription price $P = T - v$, and the consumer will be equally well off as in the base case where there is no platform whatsoever. Third, the platform can choose a freemium model, as many platforms do, where users can decide whether to pay a subscription fee to avoid ads or to enjoy for free but in the presence of ads. For those that participate in the subscription plan, the same arguments as before show that the platform will charge a subscription fee $P = T - v$ to extract full entertainment surplus from the user. On the other hand, the advertising plan will involve more aggressive advertising intensity α_N^* of the form in Proposition 5.3.1(c). The reasoning relies on Lemma 5.3.1 and the

[13] To see this, observe that if the platform were to choose an ad intensity α_N^* as in Proposition 5.3.1(c) it would drive sophisticates away. However, the platform can always attract them via a subscription plan with a sufficiently low fee, so the platform would prefer to adopt the freemium model.

fact that for advertising platforms, the participation constraint of the sophisticated agents will bind first. This means in a freemium model, the sophisticates will opt for subscription whereas the naives will opt for advertising. The participation constraint of the naives will then bind at exactly α_N^* as in Proposition 5.3.1(c).

These insights establish that the unique equilibrium is of three possibilities, two potential pooling equilibria and a potential separating equilibrium. In the two pooling equilibria (advertising-based and subscription-based), both types adopt the same plan, which is the unique plan offered by the platform. In the separating equilibrium, however, the platform offers two plans catered to each type, who self-select into the plan they prefer. Incidentally, the consumer welfare in the advertising-based plan is exactly the consumer welfare from Proposition 5.3.1(b), the consumer welfare in the subscription-based plan is exactly the base case consumer welfare, and the consumer welfare in freemium plan is exactly the consumer welfare from Proposition 5.3.1(c). As a result, consumer welfare can only *decrease* when the platform has the technology to offer a freemium model.

5.4.2 Ad Technology and Optimal Business Models

Our next result studies how the information structure of digital ads (ϕ_0 and ϕ_1) affects the business model chosen by the platform. Recall that ϕ_1 corresponds to the probability user i receives a positive signal $s = \oplus$ when indeed $\theta_i = 1$ (i.e., user i likes the product). That is, ϕ_1 can be interpreted as a measure of technological effectiveness of ads reaching their target audience. On the other hand, ϕ_0 corresponds to the probability user i receives a positive signal $s = \oplus$ when $\theta_i = 0$ (i.e., user i actually *does not* like the product). In that sense, ϕ_0 measures the ad technology's ability to manipulate consumers (in particular, naives, who wrongfully believe $\phi_0 = 0$) into believing the product would provide them with positive utility, even when it would not. Our main characterization of the platform's optimal business plan is stated next.

Theorem 5.4.1. *There exist* $0 \leq \phi_1^*(\phi_0) \leq \phi_1^{**}(\phi_0) \leq 1$ *such that*

(i) If $\phi_1 > \phi_1^{**}(\phi_0)$*, the optimal business model is* advertising-based*;*

(ii) If $\phi_1^*(\phi_0) < \phi_1 < \phi_1^{**}(\phi_0)$*, the optimal business model is* freemium*;*

(iii) If $\phi_1 < \phi_1^*(\phi_0)$*, the optimal business model is* subscription-based.

Moreover, $\phi_1^{**}(\phi_0)$ is increasing in ϕ_0 and $\phi_1^{*}(\phi_0)$ is decreasing in ϕ_0.

Theorem 5.4.1 provides a full characterization of the optimal platform model as a function of the digital ad technology. When ϕ_1 is large, there are tremendous informational gains (I_S and I_N) for both sophisticates and naives when viewing ads. This means all users are more tolerant of ads, allowing the platform to advertise more and generate additional profits for the firm. When ϕ_1 is small, ads are especially obnoxious for both sophisticates and naives, because they only provide noise. On the other hand, when ϕ_1 is an intermediate range, naives are more tolerant of ads than sophisticates, and the platform optimally separates them by offering a freemium model where sophisticates subscribe and naives do not, but are subjected to watching ads. As we pointed out, this regime is also the *worst* for consumer welfare.

The size of each of these regions depends on the parameter ϕ_0, which captures the advertiser's ability to manipulate (naive) users into believing they would like the product, even when they do not. Recall that naives have a misspecified model that $\phi_0 = 0$, and so are particularly susceptible to manipulation for larger values of ϕ_0. At the extreme where $\phi_0 = 0$ (positive signals $\oplus$ are never realized when $\theta_i = 0$), the freemium model is never optimal — both sophisticates and naives have correctly specified models, and there is no need to offer plans that separate them. At the other extreme, where $\phi_0 = \phi_1$, naives believe ads are informative but sophisticates do not, and there is maximal separation between the informational gains I_S and I_N. As ϕ_0 increases (and the manipulative ad technology develops), the platform moves from either a purely advertising-based or purely subscription-based platform to one that specifically targets naive agents with high ad intensities in a freemium-based model.

Recall that $\bar{W}^*$ denotes the average consumer welfare. The welfare implications of Theorem 5.4.1 are nuanced, as demonstrated by our next result.

Proposition 5.4.1.

(a) $\bar{W}^*$ is non-monotone in ϕ_1 for some ϕ_0;

(b) $\bar{W}^*$ is monotonically decreasing in ϕ_0 for all ϕ_1.

Improvements in the informative ad technology (ϕ_1) have non-monotonic effects on consumer welfare. As ϕ_1 increases, there are many forces at play. First, higher values of ϕ_1 correspond to stronger information about user preferences coming from digital ads, which potentially helps consumers. However, simultaneously, the platform and firm use this improved technology to then extract more consumer surplus and reel in naives. As in Theorem 5.3.1, the firm can increase its price with more effective ads and as evidenced by Theorem 5.4.1, the platform can change its business model to possibly prey more on these naive consumers. These rich interactions imply that in general $\bar{W}^*$ will not be increasing or decreasing in the informativeness of ads.

On the other hand, the technology ϕ_0 used to manipulate those who dislike the product into believing the opposite can only decrease consumer welfare. The intuition relies on Theorem 5.4.1, which shows that the platform's incentives to adopt a freemium model (with a high ad intensity that targets naives) are increasing in ϕ_0. At the same time, these targeted ads at naives are more likely to influence and more aggressive purchasing based on ad viewership. This leads to higher product prices and a decrease in consumer welfare for all types of users.

5.5 Firm-Level and Platform-Level Competition

Up until now, we have assumed there is just a single (monopolistic) firm and single (monopolistic) platform. In this section, we consider potential anti-trust solutions to remedy the reductions in welfare caused by advertising platforms. In Section 5.5.1, we study competition at the *firm level* where we assume there are multiple firms selling products who compete over advertising space on a single digital platform. In Section 5.5.2 we investigate competition at the *platform level*, where there are multiple platforms that compete for user attention and potentially a single firm's advertising business.

5.5.1 Advertising with Firm-Level Competition

There are now two product firms instead of just one.[14] Each of the firms has a single product (product 1 and product 2), and the products are completely differentiated. In particular, we assume that the user i has both $\theta_i^{(1)}$ and $\theta_i^{(2)}$ and product consumption utility:

$$\theta_i^{(1)} z_i^{(1)} + \theta_i^{(2)} z_i^{(2)} - \frac{(z_i^{(1)} + z_i^{(2)})^2}{2} - p_1 z_i^{(1)} - p_2 z_i^{(2)} .$$

[14] Our main result of this section, Proposition 5.5.1, is generalizable to N firms with N independent products, but for ease of exposition we limit our focus to just $N = 2$.

As before, $\theta_i^{(1)}$ and $\theta_i^{(2)}$ both have prior probability q of $\theta_i^{(j)} = 1$ and probability $1-q$ of $\theta_i^{(j)} = 0$ for both firms $j \in \{1, 2\}$ and are drawn independently. The platform can advertise for both firms at rates $\alpha^{(1)}$ and $\alpha^{(2)}$, with total ad intensity given by $\alpha^{(1)} + \alpha^{(2)}$. A given ad endorses product j with probability $\alpha^{(j)}/(\alpha^{(1)} + \alpha^{(2)})$, and the platform offers simultaneous contracts $(\alpha^{(1)}, m^{(1)})$ and $(\alpha^{(2)}, m^{(2)})$ to both firms. As before, we assume the probability of each ad appearing is independent across consumers and across multiple ads seen by the same consumer.

First, we generalize the user's optimal product consumption decision, taking the digital advertising on the platform as given. Each user i will have an estimate $\pi_i^{(j)}$, which is her belief that product j has $\theta_i^{(j)} = 1$, given the advertising signals she received on the platform. The user will then choose $j^* \in \arg\max_j \beta\pi_i^{(j)} - p^{(j)*}$ and consume all of product j^* as before, with $z_i^{(j^*)^*} = \beta\pi_i^{(j^*)} - p^{(j^*)*}$ and consume none of the other product (denoted $(-j^*)$), with $z_i^{(-j^*)^*} = 0$. Second, assuming all users participate on the advertising platform, firm j will solve a problem of the form,

$$\max_{p^{(j)}} \underbrace{\left(p^{(j)} - c\right)}_{\text{Profit Margin}} \cdot \underbrace{\left(Z(\alpha^{(j)}) - p^{(j)}\right)}_{\text{Aggregate Consumption}}, \tag{5.4}$$

when there is no competition, and solve a problem of the form,

$$\max_{p^{(j)}} \underbrace{\left(p^{(j)} - c\right)}_{\text{Profit Margin}} \cdot \underbrace{\left(Z(\alpha^{(j)}) - p^{(j)}\right)}_{\text{Aggregate Consumption}} \\ \cdot \left(\lambda \underbrace{\mathbb{P}\left[\pi_S^{(j)} \geq \pi_S^{(-j)} + \left(p^{(-j)} - p^{(j)}\right) \,\middle|\, \alpha^{(-j)}, \alpha^{(j)}\right]}_{\text{Probability Product } j \text{ is Preferred for Sophisticates}} + (1-\lambda) \underbrace{\mathbb{P}\left[\pi_N^{(j)} \geq \pi_N^{(-j)} + \left(p^{(-j)} - p^{(j)}\right) \,\middle|\, \alpha^{(-j)}, \alpha^{(j)}\right]}_{\text{Probability Product } j \text{ is Preferred for Naives}} \right), \tag{5.5}$$

when there is competition (where $Z(\alpha^{(j)})$ is some function of j's advertising intensity, $\alpha^{(j)}$ but not $(-j)$'s advertising intensity, $\alpha^{(-j)}$). The following result partially characterizes how this competition affects average consumer welfare relative to a monopolist seller on the platform:

Proposition 5.5.1. *The are cutoffs $0 < \underline{\lambda}^{\mathcal{F}} < \bar{\lambda}^{\mathcal{F}} < 1$ and $0 < \bar{\phi}_0^{\mathcal{F}} < \phi_1$ such that*

(i) If $\lambda > \bar{\lambda}^{\mathcal{F}}$ and $\phi_0 > \bar{\phi}_0^{\mathcal{F}}$, average consumer welfare under platform-level competition is strictly higher than under a monopolist platform;

(ii) If $\lambda < \underline{\lambda}^{\mathcal{F}}$ and $\phi_0 > \bar{\phi}_0^{\mathcal{F}}$, average consumer welfare under platform-level competition is weakly lower than under a monopolist platform.

Recall that λ captures the sophistication level of the population and ϕ_0 captures the probability of a $\oplus$ signal when $\theta_i = 0$, which also serves as a proxy for how effective digital ad targeting is on naives. Simply put, the result of Proposition 5.5.1 states that whether introducing firm-level competition helps or hurts welfare hinges on the naivety of the platform's population to be susceptible to ads. When this susceptibility is insignificant (as in (i)), consumer welfare improves because of natural competition for advertising and due to price competition across products. However, introducing a vulnerable population (as in (ii)) leads to different competitive forces, ones that use advertising as a differentiating factor, and in particular, to manipulate users into paying higher prices relative to their competitors'.

The intuition for Proposition 5.5.1(i) is straightforward. Most of the population is not heavily influenced by advertising, so firms have low willingness to pay for these ads, which is further intensified by competition for advertising spots. In this setting, platforms either switch to subscription models, or if they still use advertising, it has little influence on purchasing decisions, as most users purchase on the basis of price. All consumers end up better off with low advertising levels and lower prices after a second firm enters the market.

However, product competition unfolds quite differently in the setting of Proposition 5.5.1(ii). The intuition is best seen by considering a single ad shown to each naive agent, which happens to be either firm 1's or firm 2's product (with equal probability). There is a high likelihood the naive agent is convinced by the ad, and is willing to pay a premium for the product she saw advertised over her competitor's. If the agent saw firm 1's advertisement, firm 1 knows it could win that agent's business even if its price was higher than its competitor's. On the other hand, if the naive agent saw firm 2's ad, she would require a *discount* to purchase firm 1's product over firm 2's. As a consequence, firm 1 should price agents who saw firm 2's ad out of the market altogether, which leads both firm 1 and firm 2 to have higher prices than either were a monopolist. This leads to a reduction in consumer welfare by stronger incentives to adopt the freemium models of Section 5.4 that prey on the naives in this way.

5.5.2 Platform-Level Competition

There are now two platforms instead of just one (with a single firm).[15] Both platforms simultaneously set subscription prices P_1 and P_2 and offer contracts to the firm, (α_1, m_1) and (α_2, m_2). At $t = 2$, the firm will choose the better of the two contracts,[16] then users will choose how to allocate their time in $t = 3$,[17] with the rest of the game played out in $t = 4$ and $t = 5$ as described in Section 5.2. Similar to Proposition 5.5.1, we find that:

Proposition 5.5.2. *The are cutoffs* $0 < \underline{\lambda}^{\mathcal{P}} < \bar{\lambda}^{\mathcal{P}} < 1$ *and* $0 < \bar{\phi}_0^{\mathcal{P}} < \phi_1$ *such that*

(i) If $\lambda > \bar{\lambda}^{\mathcal{P}}$ *and* $\phi_0 > \bar{\phi}_0^{\mathcal{P}}$*, average consumer welfare under platform-level competition is strictly higher than under a monopolist platform;*

(ii) If $\lambda < \underline{\lambda}^{\mathcal{P}}$ *and* $\phi_0 > \bar{\phi}_0^{\mathcal{P}}$*, average consumer welfare under platform-level competition is weakly lower than under a monopolist platform.*

Proposition 5.5.2 is the analog to Proposition 5.5.1, although the competitive forces at play are slightly different, despite yielding similar conclusions. Because the two platforms are identical (as opposed to firm-level competition with differentiated products), Bertrand competition effects are much stronger. Platforms cannot extract surplus from firms in equilibrium, and thus offer contracts with $m = 0$ and α chosen strategically to best benefit the firm. Similarly, platform subscribers will look for the cheaper of the two offered plans, so when some users do subscribe, the subscription fee will always be $P = 0$. This simplification allows us to more easily to characterize the equilibria.

In regime (i), manipulation via targeted digital advertising is not as possible because the population is mostly sophisticated. This will naturally result in more subscription-based users, but the platform-level competition will drive subscriptions prices down and these users will experience surplus from entertainment. While advertisement-based plans might still be available for the small fraction of naive users, average consumer welfare will still rise and naives

[15] Like with firm-level competition (as remarked in Footnote 14), platform-level competition with more than two platforms does not materially affect our analysis.

[16] The firm could also technically accept *both* contracts, but observe this is weakly dominated. If both platforms offer the same ad intensity α, the firm should accept the cheaper of the two. If the platforms offer different ad intensities, say, $\alpha_1 < \alpha_2$, then all ad-based users will go to platform 1 so it is an equally good response to only accept firm 1's contract.

[17] Also, observe that the user would never prefer to "split" time between the two platforms except in knife-edge cases where she gets equal utility from allocating her time to both. As a convention, we suppose that if the user is perfectly indifferent then chooses one platform at random.

may even be better off. This may seem surprising in juxtaposition with Section 5.4, where freemium business models were classified as minimizing consumer welfare. However, in the case of Proposition 5.5.2(i), the freemium business model arises out of competition, and can overall *increase* welfare because it lowers subscription fees and simultaneously forces the firm to set reasonable prices, not exploit naives' advertising susceptibilities.

In regime (ii), the same freemium business model that comes out of platform competition leads to reduced average consumer welfare. The reason is that sophisticated users are generally more inclined to switch to subscription-based plans than naives are. A monopolist platform which has more power to set the business model could force all users to engage with advertising. However, with competition from another platform, it can start to offer subscriptions that will steal market share from the original platform, which will only exploit naives on their platform. Because this competition is unavoidable, it might as well increase the intensity of advertising knowing it will drive sophisticates anyway. Overall, in a population largely consisting of naives, this reduces average consumer welfare.

Propositions 5.5.1 and 5.5.2, taken together, show that anti-trust regulation that involves breaking up firms or platforms is an imperfect solution to the digital advertising problem at best. In naive populations, where their welfare is already substantially hurt by the business models chosen by platforms, competition at both the firm-level and platform-level will exacerbate the problem.

5.6 Policy: Digital Ad Taxation

Finally, in light of the anti-trust analysis of Section 5.5, we propose a new suggestion, a digital advertising tax. For this, we suppose that the platform is currently monetized only through digital advertising in equilibrium (despite being able to offer subscription services or a mixed freemium model).[18] As observed in Section 5.4, this leads to worse consumer welfare than if the platform were non-existent or offered subscriptions (with no digital ads) to all users, but leads to better welfare compared to the freemium business model.

In an attempt to correct consumer welfare back to base-case levels, we introduce a digital

[18] This is the most interesting case to analyze so we focus on it here. But the result of Theorem 5.6.1 immediately extends to the freemium business model, except where the low tax rate is not effective (there is no middle region). If the current business model is already subscription, then an ad tax is meaningless.

ad tax that charges the platform a tax rate ζ on all revenue earned via digital ads.[19] In particular, the transfer to the platform with an accepted contract (α, m) is given by $(1-\zeta)m$, after the tax is levied. Our next result characterizes the effect of such a taxation policy.

Theorem 5.6.1. *There exist* $0 < \underline{\zeta} < \overline{\zeta} < 1$ *such that*

(a) If $\zeta \in (0, \underline{\zeta})$, $\bar{W}^*$ *remains unaffected after the taxation policy;*

(b) If $\zeta \in (\underline{\zeta}, \overline{\zeta})$, $\bar{W}^*$ *decreases after the taxation policy;*

(c) If $\zeta \in (\overline{\zeta}, 1)$, $\bar{W}^*$ *increases to base-case welfare after the taxation policy.*

Consumer welfare is non-monotone in the level of taxation, and in particular, an ad tax can be ineffective or even damaging if not implemented with sufficient aggression. The intuition for Theorem 5.6.1 echoes similar non-monotonicity of Proposition 5.4.1(a). When the ad tax rate is small, the tax does not adequately deter the platform from generating revenue purely based on advertising. Once the tax becomes sizable, the platform does not switch to fully relying on subscription but instead switches to the freemium business model of Theorem 5.4.1(ii). Under this model, the platform intensifies advertising among a particularly susceptible community, and simply charges others for subscription. As we have remarked in Section 5.4, this depresses consumer welfare to its worst levels. However, once the tax rate becomes sufficiently high, the platform is forced to switch to a purely subscription-based model as in Proposition 5.4.1(iii). This has the intended consequence of restoring consumer welfare back to base-case levels.

Remark — Theorem 5.6.1 provides a cautionary tale for implementing a digital ad tax to incentivize the platform to substitute away from ad-based revenue sources. The result also provides the basis for a *progressive* ad tax, with a marginal tax rate $\zeta(m)$ that is increasing in m to avoid region (b) in Theorem 5.6.1. Such a progressive tax can be constructed by ensuring that $\frac{1}{m_F}\int_0^{m_F} \zeta(m)\, dm \geq \bar{\zeta}$,[20] but where $\zeta(0) = 0$. Such a tax can be more effective at ensuring that platforms with little revenue (of any kind) are not harshly punished with an aggressive tax rate $\zeta > \bar{\zeta}$.

[19] Such an advertising tax has been proposed recently by economist and entrepreneur Paul Romer, as way to shift platform incentives toward subscription-based revenue instead of ad-based revenue (see https://adtax.paulromer.net/).

[20] Here, m_F is the equilibrium transfer from the firm to the platform under the freemium model, which is independent of the tax conditional on being in regime (b) of Theorem 5.6.1.

Appendix A

Misinformation: Strategic Models

A.1 A Model of Online Misinformation

A.1.1 Proofs

Auxiliary Lemmas

We define a (mixed-strategy) strategy σ_i for agent i to be a map from priors b_i to elements of the simplex $\Delta(\{\mathcal{D}, \mathcal{I}, \mathcal{S}\})$. In others words, σ_i specifies for each ideological prior b_i of agent i the probability that she will play each of the three actions, $\mathcal{D}$, $\mathcal{I}$, and $\mathcal{S}$. We let $\boldsymbol{\sigma}_{-i}$ denote the (vector of) strategies of all agents other than agent i.

Lemma A.1.1. *Given any set of strategies $\boldsymbol{\sigma}_{-i}$, agent i's best response is a cutoff strategy with cutoffs (b_i^*, b_i^{**}) such that if $b_i < b_i^*$ agent i dislikes ($\mathcal{D}$), if $b_i^* < b_i < b_i^{**}$ agent i ignores ($\mathcal{I}$), and if $b_i > b_i^*$ agent i shares ($\mathcal{S}$).*

Proof of Lemma A.1.1. When agent i receives an article, she forms (ex-post) belief π_i about the article's veracity which depends only on the observables (r, m). By Bayes' rule:

$$\pi_i \equiv \mathbb{P}[\nu = \mathcal{T} \,|, \, r, m = R] = \frac{\mathbb{P}[m = R \,|\, r, \nu = \mathcal{T}]\mathbb{P}[\nu = \mathcal{T} \,|\, r]}{\mathbb{P}[m = R \,|\, r, \nu = \mathcal{M}]\mathbb{P}[\nu = \mathcal{M} \,|\, r] + \mathbb{P}[m = R \,|\, r, \nu = \mathcal{T}]\mathbb{P}[\nu = \mathcal{T} \,|\, r]} \,.$$

By the law of total probability, we have:

$$\begin{aligned}\mathbb{P}[m = R \,|\, r, \nu = \mathcal{T}] = \mathbb{P}[m = R \,|\, \nu = \mathcal{T}] &= \mathbb{P}[m = R \,|\, \theta = R, \nu = \mathcal{T}]\mathbb{P}[\theta = R] \\ +\mathbb{P}[m = R \,|\, \theta = L, \nu = \mathcal{T}]\mathbb{P}[\theta = L] & \\ &= pb_i + (1-p)(1-b_i)\ ; \\ \mathbb{P}[m = R \,|\, r, \nu = \mathcal{M}] = \mathbb{P}[m = R \,|\, \nu = \mathcal{M}] &= \mathbb{P}[m = R \,|\, \theta = R, \nu = \mathcal{M}]\mathbb{P}[\theta = R] \\ +\mathbb{P}[m = R \,|\, \theta = L, \nu = \mathcal{M}]\mathbb{P}[\theta = L] & \\ &= qb_i + (1-q)(1-b_i)\ .\end{aligned}$$

Putting these together we obtain equation (2.2). Moreover, π_i is monotone in b_i since

$$\frac{\partial \pi_i}{\partial b_i} = \frac{(1-\phi(r))\phi(r)(p-q)}{(1-b_i+q(1-\phi(r))(2b_i-1)-p(\phi(r)-2\phi(r)b_i))^2} > 0.$$

Note that $U_i(\mathcal{I})$ and $U_i(\mathcal{D})$ is independent of $\boldsymbol{\sigma}_{-i}$, and in particular $\mathcal{D}$ is a better response to $\mathcal{I}$ if and only if $\pi_i < (\tilde{u}-\tilde{c})/\tilde{u}$. Because π_i is monotone in b_i, this implies there exists some $\tilde{b}$ where $\mathcal{D}$ is a better response to $\mathcal{I}$ if and only if $b_i < \tilde{b}$ (where $\tilde{b} = 1$ if disliking dominates ignoring and $\tilde{b} = 0$ if ignoring dominates disliking). Next, recall that the payoff to sharing is $U_i(\mathcal{S}) = U_i^{(1)} + U_i^{(2)}$, where $U_i^{(1)} = u\mathbf{1}_{\nu=\mathcal{T}} - c\mathbf{1}_{\nu=\mathcal{M}}$ and $U_i^{(2)} = \kappa S_i - dD_i$. Observe that, as before, $U_i^{(1)}$ is independent of $\boldsymbol{\sigma}_{-i}$ and has expected payoff $(u+c)\pi_i - c$, which is monotonically increasing in π_i. Moreover, $\mathbb{E}_{\mathbf{P},\boldsymbol{\sigma}_{-i}}[\kappa S_i - dD_i]$ does not depend on b_i. Because π_i is monotone in b_i, we see that $U_i(\mathcal{S})$ is increasing in b_i, $U_i(\mathcal{I})$ is constant in b_i (it is always zero), and $U_i(\mathcal{D})$ is decreasing in b_i (it is equal to $\tilde{u}(1-\pi_i)-\tilde{c}$). This implies that either (i) ignoring dominates sharing, (ii) sharing dominates ignoring, or (iii) $U_i(\mathcal{S}) = 0$ for some prior b':

(i) If ignoring dominates sharing, we set $(b_i^*, b_i^{**}) = (\tilde{b}, 1)$.

(ii) If sharing dominates ignoring, then either sharing dominates disliking (in which case set $(b_i^*, b_i^*) = (0,0)$), disliking dominates sharing (in which case we set $(b_i^*, b_i^{**}) = (1,1)$), or there exists some prior b'' where $U_i(\mathcal{S}) = U_i(\mathcal{D})$ (in which case set $(b_i^*, b_i^{**}) = (b'', b'')$).

(iii) Otherwise, if $\tilde{b} < b'$, set $(b_i^*, b_i^{**}) = (\tilde{b}, b')$; however, if $\tilde{b} \geq b'$, then we set $(b_i^*, b_i^{**}) = (b', b')$.

This is of the cutoff form claimed in the lemma. ∎

An immediate consequence of Lemma A.1.1 is that any Bayesian-Nash equilibrium must be in cutoff strategies for all agents. Hence, we can limit our attention to cutoff strategies (b_i^*, b_i^{**})

for every agent i, which can be represented as $(\mathbf{b}^*, \mathbf{b}^{**})$ in vector notation. This is a partially-ordered set according to the component-wise order $\succeq$. Hence, the cutoff space $\mathbf{B} = [0,1]^{2N}$ forms a *complete lattice.*[1]

Next, we define a map $\psi : \mathbf{B} \to \mathbf{B}$ that maps cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$ to best-response cutoffs $(\mathbf{b}^{*,BR}, \mathbf{b}^{**,BR})$. This map is well-defined because (i) H is a continuous distribution, so we need not specify the strategies of agents precisely on the cutoffs, and (ii) by Lemma A.1.1, for any set of strategies $\boldsymbol{\sigma}_{-i}$ (including the cutoff strategies given by $(\mathbf{b}^*, \mathbf{b}^{**})$), all agents' best responses are in cutoff form.

Lemma A.1.2. *The map ψ preserves the component-wise order $\succeq$.*

Proof of Lemma A.1.2. Consider some $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**}) \succeq (\mathbf{b}^*, \mathbf{b}^{**})$. Fixing an article with observables (r, m), $U_i(\mathcal{D})$, $U_i(\mathcal{I})$ and $U_i^{(1)}$ are independent of $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$ and $(\mathbf{b}^*, \mathbf{b}^{**})$. However, for $U_i^{(2)}$ we have:

$$\begin{aligned}\mathbb{E}_{\mathbf{P},(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}[\kappa S_i - dD_i] &= \sum_{j=1}^{N} p_{ij} \left(\kappa \mathbb{P}_{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}[a_j = \mathcal{S}] - d\mathbb{P}_{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}[a_j = \mathcal{D}]\right) = \sum_{j=1}^{N} p_{ij} \left(\kappa \mathbb{P}_H[b_j > \hat{b}_j^{**}] - d\mathbb{P}_H[b_j < \right. \\ &\leq \sum_{j=1}^{N} p_{ij} \left(\kappa \mathbb{P}_H[b_j > b_j^{**}] - d\mathbb{P}_H[b_j < b_j^*]\right) = \mathbb{E}_{\mathbf{P},(\tilde{\mathbf{b}}^*,\tilde{\mathbf{b}}^{**})}[\kappa S_i - dD_i]\,.\end{aligned}$$

As a result, $U_i^{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}(\mathcal{S}) \leq U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S})$. As in Lemma A.1.1, we define $\tilde{b}$ as the prior where $U_i(\mathcal{D}) = 0$ if such a $\tilde{b}$ exists, otherwise let $\tilde{b} = 0$ if ignoring dominates disliking and $\tilde{b} = 1$ if disliking dominates ignoring. Observe that $\tilde{b}$ is the same for both $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$ and $(\mathbf{b}^*, \mathbf{b}^{**})$. We have three cases for the best-response cutoffs $(\mathbf{b}^{*,BR}, \mathbf{b}^{**,BR})$ given other agents' cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$ (which we compare to $(\hat{\mathbf{b}}^{*,BR}, \hat{\mathbf{b}}^{**,BR})$ given other agents' cutoffs $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$):

(i) Ignoring dominates sharing for agent i (for given cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$). Then by virtue of $U_i^{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}(\mathcal{S}) \leq U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S})$, ignoring dominates sharing with $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$ as well. Thus, $(b_i^{*,BR}, b_i^{**,BR}) = (\hat{b}_i^{*,BR}, \hat{b}_i^{**,BR}) = (\tilde{b}, 1)$.

(ii) Sharing dominates ignoring for agent i (for given cutoffs $(\mathbf{b}^*, \mathbf{b}^{**})$). Then either sharing dominates disliking (in which case $(b_i^{*,BR}, b_i^{**,BR}) = (0, 0) \preceq (\hat{b}_i^{*,BR}, \hat{b}_i^{**,BR})$ trivially), or there exists some prior b'' where $U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) = U_i(\mathcal{D})$ denoted by b'' and $(b_i^{*,BR}, b_i^{**,BR}) = (b'', b'')$.

[1] Note that for any collection of cutoffs $\{(\mathbf{b}^{*,(1)}, \mathbf{b}^{**,(1)}), (\mathbf{b}^{*,(2)}, \mathbf{b}^{**,(2)}), \ldots, \}$ in the cutoff space, there is a greatest lower bound given by the component-wise infimum and a least upper bound given by the component-wise supremum.

Moreover, because $U_i^{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}(\mathcal{S}) \leq U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) = U_i(\mathcal{D})$ at prior b'', for an agent with prior b'', playing $\mathcal{D}$ is a (weakly) better response than sharing when other agents play according to cutoffs $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$. By monotonicity of $U_i(\mathcal{S})$ and $U_i(\mathcal{D})$ in prior b_i, this implies that $b_i^{**,BR} \leq \hat{b}_i^{**,BR}$. If ignoring is never a best response when $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$, then $\hat{b}_i^{*,BR} = \hat{b}_i^{**,BR}$. Otherwise, $\hat{b}_i^{*,BR} = \tilde{b} \geq b_i^{**,BR} = b_i^{*,BR}$.

(iii) $U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) = 0$ for some prior b' for agent i. Then $U_i^{(\hat{\mathbf{b}}^*,\hat{\mathbf{b}}^{**})}(\mathcal{S}) \leq U_i^{(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) = 0$ implies that for an agent with prior b' playing $\mathcal{I}$ is a (weakly) better response than sharing when other agents play according to $(\hat{\mathbf{b}}^*, \hat{\mathbf{b}}^{**})$. By monotonicity of $U_i(\mathcal{S})$ in prior b_i, this implies that $b_i^{**,BR} \leq \hat{b}_i^{**,BR}$. If $\tilde{b} < b'$, then $b_i^{*,BR} = \hat{b}_i^{*,BR} = \tilde{b}$; otherwise, if $\tilde{b} \geq b'$, $b_i^{*,BR} = b_i^{**,BR} = b' \leq \hat{b}_i^{*,BR}$.

This establishes that $(\hat{\mathbf{b}}_i^{*,BR}, \hat{\mathbf{b}}_i^{**,BR}) \succeq (\mathbf{b}_i^{*,BR}, \mathbf{b}_i^{**,BR})$, so the order $\succeq$ is preserved by ψ. ∎

Lemma A.1.3. *An increase in polarization of beliefs can be constructed via the following process: take every belief b_i and either (i) add some $\epsilon_i > 0$ to b_i if $b_i > 1/2$, or (ii) subtract some $\epsilon_i > 0$ to b_i if $b_i < 1/2$.*

Proof of Lemma A.1.3. Let H_2 be more polarized than H_1. For part (i), note that $H_1(b_i^1) = \alpha > 1/2$, so by single-crossing at $H_1^{-1}(1/2) = H_2^{-1}(1/2)$, we know that $H_2^{-1}(\alpha) - H_1^{-1}(\alpha) > 0$. Thus, for some $b_i^2 > b_i^1$, we have $H_2^{-1}(\alpha) = b_i^2$, or in other words, $H_2(b_i^2) = \alpha$. Setting $\epsilon_i = b_i^2 - b_i^1 > 0$ in this fashion for all $b_i > 1/2$ accomplishes claim (i). For part (ii), note that $H_1(b_i^1) = \alpha < 1/2$, so by single-crossing at $H_1^{-1}(1/2) = H_2^{-1}(1/2)$, we know that $H_2^{-1}(\alpha) - H_1^{-1}(\alpha) < 0$. Thus, for some $b_i^2 < b_i^1$, we have $H_2^{-1}(\alpha_1) = b_i^2$, or in other words, $H_2(b_i^2) = \alpha$. Setting $\epsilon_i = b_i^1 - b_i^2 > 0$ in this fashion for all $b_i < 1/2$ accomplishes claim (ii). ∎

Lemma A.1.4. *If $\kappa \leq \bar{\kappa} \equiv (c\tilde{c} - u(\tilde{u} - \tilde{c}))/(\tilde{u}N)$, then for any agent i:*

(i) If $b_i^* > 0$ and $b_i^{**} < 1$, then $b_i^{**} > b_i^*$;

(ii) For all $\bar{b} < 1$, there exists $\tilde{r} > 0$ such that agent i plays $\mathcal{D}$ in any equilibrium for an article with $r < \tilde{r}$ and on any sharing network $\mathbf{P}$, provided that $b_i < \bar{b}$.

Proof of Lemma A.1.4. For part (i), by way of contradiction suppose that $b_i^* = b_i^{**}$. Then for an agent with prior b_i^* (and corresponding ex-post belief π_i^* that the article is truthful), it must be the case that:

$$\tilde{u}(1-\pi_i) - \tilde{c} = u\pi_i - c(1-\pi_i) + \mathbb{E}[\kappa S_i - dD_i] \geq 0\,.$$

Re-arranging we get that $\pi_i = \frac{\tilde{u}-\tilde{c}+c-\mathbb{E}[\kappa S_i - dD_i]}{\tilde{u}+u+c}$. Substituting into the payoff for action $\mathcal{D}$, we see that:

$$U_i(\mathcal{D}) = \tilde{u}\left(\frac{u+\tilde{c}+\mathbb{E}[\kappa S_i - dD_i]}{\tilde{u}+u+c}\right) - \tilde{c} \leq \tilde{u}\left(\frac{u+\tilde{c}+\kappa N}{\tilde{u}+u+c}\right) - \tilde{c} < \tilde{u}\left(\frac{u+\tilde{c}+\bar{\kappa} N}{\tilde{u}+u+c}\right) - \tilde{c} \leq 0 \,.$$

By assumption, $U_i(\mathcal{S}) = U_i(\mathcal{D}) < 0$, but since $U_i(\mathcal{I}) = 0$, ignoring is the best response at prior b_i^*, which is a contradiction.

For part (ii), notice by equation (2.2), for a fixed $b < 1$, as $r \to 0$, $\pi_i \to 0$, and therefore:

$$U_i(\mathcal{S}) = u\pi_i - c(1-\pi_i) + \mathbb{E}[\kappa S_i - dD_i] < u\pi_i - c(1-\pi_i) + \bar{\kappa} N \leq u\pi_i - c(1-\pi_i) + \frac{c}{N} N \overset{r\to 0}{=} -c + c = 0 \,.$$

where the last inequality follows from the observation that:

$$\bar{\kappa} \equiv \frac{c\tilde{c} - u(\tilde{u}-\tilde{c})}{\tilde{u}N} < \frac{c\tilde{c}}{\tilde{u}N} < \frac{c}{N} \,,$$

because $\tilde{u} > \tilde{c}$. Thus, as $r \to 0$, ignoring is a better response than sharing. But note that $U_i(\mathcal{D}) = \tilde{u}(1-\pi_i) - \tilde{c} \overset{r\to 0}{=} \tilde{u} - \tilde{c} > 0$, so as $r \to 0$, disliking is a better response than ignoring. As a result, disliking is a best response for any fixed $b < 1$ as $r \to 0$. The claim in (ii) thus follows from continuity of equation (2.2). ∎

Proofs from Section 3

Proof of Theorem 2.1.1. Claim (ii) follows directly from Lemma A.1.1 and establishes that the Bayesian-Nash equilibria are the fixed points of the map ψ. Clearly the cutoff space B is convex and compact (it is defined by $[0,1]^{2N}$). To see that ψ is continuous, notice that for $\psi : (\mathbf{b}^*, \mathbf{b}^{**}) \mapsto (\mathbf{b}^{*,BR}, \mathbf{b}^{**,BR})$, $\mathbb{E}_{\mathbf{P},(\mathbf{b}^*,\mathbf{b}^{**})}[U_i^{(2)}]$ is continuous because H is continuous (and $U_i(\mathcal{D})$, $U_i(\mathcal{I})$, and $U_i^{(1)}$ do not depend on $(\mathbf{b}^*, \mathbf{b}^{**})$). Moreover, by the same reasoning as in Lemma A.1.2, $U_i^{\mathbf{P},(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S})$ and $U_i^{\mathbf{P},(\mathbf{b}^*,\mathbf{b}^{**})}(\mathcal{S}) - U_i(\mathcal{D})$ are monotone and continuous. Because these expressions are continuous in $(\mathbf{b}^*, \mathbf{b}^{**})$, the corresponding best-response cutoffs, $(\mathbf{b}^{*,BR}, \mathbf{b}^{**,BR})$ are also continuous in $(\mathbf{b}^*, \mathbf{b}^{**})$. By Brouwer's fixed-point theorem, there exists a Bayesian-Nash equilibrium, proving (i).

Finally, noting that the cutoff space B is a complete lattice and ψ preserves the component-wise order $\succeq$ (by Lemma A.1.2), Tarski's fixed-point theorem establishes that the set of equilibrium cutoffs forms a lattice (see Tarski (1955)). By definition of a lattice order, there exists a least-sharing equilibrium (largest b**) and a most-sharing equilibrium (smallest b**). ∎

Proof of Proposition 2.1.1. Recall that π_i is given by equation (2.2) and provides the (ex-post) belief of the article's veracity conditional on observables (r, m). Also observe that:

$$\frac{\partial \pi_i}{\partial r} = \frac{(1 - b_i + p(2b_i - 1))(1 - b_i + q(2b_i - 1))}{(1 - b_i + q(1 - \phi(r))(2b_i - 1) - p(\phi(r) - 2\phi(r)b_i))^2}\phi'(r) \, .$$

Because $\phi'(r) > 0$, it is clear that when $b_i > 1/2$, $\partial\pi_i/\partial r > 0$. When $b_i < 1/2$, $1 - b_i + p(2b_i - 1)$ is minimized when $p = 1$, in which case it is equal to $b_i \geq 0$ (and with this inequality strict whenever $p < 1$). Similarly, when $b_i < 1/2$, $1 - b_i + q(2b_i - 1)$ is minimized when $q = 1/2$, in which case it is equal to $1/2 > 0$. Thus, $\partial\pi_i/\partial r > 0$ for all b_i.

Similarly, when $\phi' \geq \phi$, a reliability score r with misinformation structure ϕ' can be translated into a higher reliability score $r' \geq r$ under misinformation structure ϕ (because both ϕ, ϕ' are monotonically increasing). As a consequence, a decrease in misinformation is isomorphic to greater reliability of the articles. It is thus sufficient to prove the latter leads to uniformly more sharing in both the least and the most sharing equilibria.

Note that the social media game is supermodular and has increasing differences in reputability. To see this, note that for all $r' \geq r$:

$$[U_i(\mathcal{S}, r') - U_i(\mathcal{I}, r')] - [U_i(\mathcal{S}, r) - U_i(\mathcal{I}, r)] = U_i(\mathcal{S}, r') - U_i(\mathcal{S}, r) = U_i^{(1)}(r') - U_i^{(1)}(r) = (u + c)(\pi_i(r') - \pi_i(r)$$

which is non-negative via the above observation that $\frac{\partial \pi_i}{\partial r} > 0$. Similarly, for all $r' \geq r$:

$$\begin{aligned}[U_i(\mathcal{S}, r') - U_i(\mathcal{D}, r')] - [U_i(\mathcal{S}, r) - U_i(\mathcal{D}, r)] &= [U_i(\mathcal{S}, r') - U_i(\mathcal{S}, r)] + [U_i(\mathcal{D}, r') - U_i(\mathcal{D}, r)] \\ &= (u + c)(\pi_i(r') - \pi_i(r)) + \tilde{u}(\pi_i(r') - \pi_i(r)) \, ,\end{aligned}$$

which is non-negative via the same observation. Finally, for all $r' \geq r$:

$$[U_i(\mathcal{I}, r') - U_i(\mathcal{D}, r')] - [U_i(\mathcal{I}, r) - U_i(\mathcal{D}, r)] = U_i(\mathcal{D}, r') - U_i(\mathcal{D}, r) = \tilde{u}(\pi_i(r') - \pi_i(r)) \, ,$$

which, again, is non-negative. Thus, via Topkis's monotone comparative statics theorem (see Topkis (1998)), there is uniformly more sharing.

Similarly, the social media game is supermodular and has increasing differences in sensationalism and (the negative of) reputational concerns. To see this, note for all $\kappa' \geq \kappa$ and $d' \leq d$:

$$[U_i(\mathcal{S}, \kappa', d') - U_i(\mathcal{I}, \kappa', d')] - [U_i(\mathcal{S}, \kappa, d) - U_i(\mathcal{I}, \kappa, d)] = U_i^{(2)}(\kappa', d') - U_i^{(2)}(\kappa', d') = (\kappa' - \kappa)S_i + (d - d')D_i$$

which is non-negative. Moreover, note that comparing $\mathcal{S}$ and $\mathcal{D}$ is identical to comparing $\mathcal{S}$ and $\mathcal{I}$ because parameters (κ, d) affect both $\mathcal{I}$ and $\mathcal{D}$ identically (they only factor into the payoff of action $\mathcal{S}$). For this same reason, we note that $[U_i(\mathcal{I}, \kappa', d') - U_i(\mathcal{D}, \kappa', d')] - [U_i(\mathcal{I}, \kappa, d) - U_i(\mathcal{D}, \kappa, d)] = 0$. Thus, via Topkis's theorem, there is uniformly more sharing. ▮

Proofs from Section 4

Proof of Lemma 2.1.1. To obtain a contradiction, suppose that there exists an agent i and an agent j with $\ell_i = \ell_j$ but either (i) $b_i^* \neq b_j^*$ or (ii) $b_i^{**} \neq b_j^{**}$.

Without loss of generality, suppose that $b_i^* < b_j^*$. By way of contradiction suppose $b_i^{**} > b_i^*$, and consider priors $\tilde{b} \in (b_i^*, \min\{b_j^*, b_i^{**}\})$ where agent i would ignore but agent j with that same prior would dislike. However, both agents with prior $\tilde{b}$ receive payoff $\tilde{u}(1 - \pi(\tilde{b})) - \tilde{c}$ from disliking and payoff of 0 from ignoring. Thus, one of them must not be playing a best response. This establishes that $b_i^{**} = b_i^*$.

Thus, when agents i and j both have some prior $b' \in (b_i^*, b_j^*)$, agent i shares and agent j dislikes. By symmetry of agent i and j's network positions, it is clear that for agent i and agent j with prior b' that $U_j(\mathcal{S}) - U_i(\mathcal{S}) = p_s(\kappa + d)$. Similarly, $U_j(\mathcal{D}) - U_i(\mathcal{D}) = 0$. But in this case,

$$[U_j(\mathcal{S}) - U_i(\mathcal{S})] - [U_j(\mathcal{D}) - U_i(\mathcal{D})] = [U_j(\mathcal{S}) - U_j(\mathcal{D})] + [U_i(\mathcal{D}) - U_i(\mathcal{S})] = p_s(\kappa + d) > 0\,.$$

This implies that either $[U_j(\mathcal{S}) - U_j(\mathcal{D})] > 0$ or $[U_i(\mathcal{D}) - U_i(\mathcal{S})] > 0$ (or both). This yields a contradiction because at prior b', it is supposed to be a best response for agent j to play $\mathcal{D}$ and a best response for agent i to play $\mathcal{S}$. Thus, $b_i^* = b_j^*$.

Without loss of generality, suppose that $b_i^{**} < b_j^{**}$. If $b_i^{**} \leq b_j^*$, then for priors $b'' \in (b_i^{**}, b_j^*)$, agent i shares and agent j dislikes. Via the same reasoning as in the previous paragraph, this is a contradiction, so $b_j^* < b_i^{**} < b_j^{**}$. Let us consider some prior $\hat{b} \in (b_i^{**}, b_j^{**})$, where agent i shares and agent j ignores. By symmetry of agent i and j's network positions, it is clear that for agent i

and agent j with prior $\hat{b}$ that $U_j(\mathcal{S}) - U_i(\mathcal{S}) = p_s\kappa$. Similarly, $U_j(\mathcal{I}) - U_i(\mathcal{I}) = 0$. Then notice that:

$$[U_j(\mathcal{S}) - U_i(\mathcal{S})] - [U_j(\mathcal{I}) - U_i(\mathcal{I})] = [U_j(\mathcal{S}) - U_j(\mathcal{I})] + [U_i(\mathcal{I}) - U_i(\mathcal{S})] = p_s\kappa > 0\,.$$

This implies that either $[U_j(\mathcal{S}) - U_j(\mathcal{I})] > 0$ or $[U_i(\mathcal{I}) - U_i(\mathcal{S})] > 0$ (or both). However, this is a contradiction because at prior b', it is supposed to be a best response for agent j to play $\mathcal{I}$ and a best response for agent i to play $\mathcal{S}$. Thus, $b_i^{**} = b_j^{**}$. ∎

Proof of Theorem 2.1.2. For part (a), let us consider belief $b^{(2)} < 1$ and a reliability threshold $\underline{r}$ such that for all $\mathbf{P}$, all agents with $b < b^{(2)}$ choose $\mathcal{D}$ in every equilibrium (including the most-sharing equilibrium) whenever the article has reliability $r < \underline{r}$. Such an $\underline{r}$ exists by Lemma A.1.4(ii). Thus, for all $r < \underline{r}$, every agent on an island $\ell \geq 2$ dislikes in the most-sharing equilibrium, regardless of $\mathbf{P}$.

Next, we consider an increase in homophily (while holding expected degree fixed). By our choice of $\underline{r}$, all agents on islands $\ell \geq 2$ still dislike in the most-sharing equilibrium whenever $r < \underline{r}$. We can thus consider the social media game that only involves island 1, treating islands 2 through k as automata that always dislike. Before the shift in homophily, consider the equilibrium cutoffs (b_1^*, b_1^{**}) for island 1 in the most-sharing equilibrium (the same for all agents on island 1, per Lermma 2.1.1) and let $\mathbf{B}_1$ denote the modified cutoff space defined by all cutoffs $(\hat{b}_1^*, \hat{b}_1^{**}) \preceq (b_1^*, b_1^{**})$. Finally we define a map $\varphi : \mathbf{B}_1 \to \mathbf{B}_1$ that maps cutoffs in $\mathbf{B}_1$, $(\hat{b}_1^*, \hat{b}_1^{**})$, to best-response cutoffs $(\hat{b}_1^{*,BR}, \hat{b}_1^{**,BR})$, given that agents on island 1 play according $(\hat{b}_1^*, \hat{b}_1^{**})$. By the arguments in Lemma A.1.2, φ preserves $\succeq$ and $\mathbf{B}_1$ is a complete sublattice, provided that the map φ is well-defined in that it always maps to an element in $\mathbf{B}_1$.

To establish this, consider the utility $U_1(\mathcal{S})$ of sharing on island 1 with homophily parameters (p_s, p_d), holding fixed the cutoff strategy $(\hat{b}_1^*, \hat{b}_1^{**})$ and the expected degree of each agent on island 1, ζ_1. Thus, we can write $p_d = (\zeta - N_1 p_s)/(N - N_1)$ and observe then that

$$U_1(\mathcal{S}) = U_1^{(1)} + \kappa N_1 p_s (1 - H(\hat{b}_1^{**})) - d\left(N_1 p_s H(\hat{b}_1^*) + \frac{\zeta - N_1 p_s}{N - N_1} \cdot (N - N_1)\right)\,,$$

and in particular, $\partial U_1(\mathcal{S})/\partial p_s = \kappa N_1(1 - H(\hat{b}_1^{**})) + d N_1 (1 - H(\hat{b}^*)) > 0$. Therefore, if we compare utility $U_1'(\mathcal{S})$ after the increase in homophily to $U_1(\mathcal{S})$ before the increase in homophily (leaving $(\hat{b}_1, \hat{b}_1^{**})$ fixed), we see that $U_1'(\mathcal{S}) \geq U_1(\mathcal{S})$. Hence, φ necessarily maps any cutoffs in $\mathbf{B}_1$ into $\mathbf{B}_1$.

Applying Again Tarski's fixed-point theorem, the set of fixed points (and thus Bayesian-Nash equilibria) form a lattice within the space of cutoffs $\mathbf{B}_1$. Moreover, there is a most-sharing equilibrium in $\mathbf{B}_1$, which is also the most-sharing equilibrium in $\mathbf{B}$. We denote this equilibrium by $(b_1^{*'}, b_1^{**'})$ and note that $(b_1^{*'}, b_1^{**'}) \preceq (b_1^{*}, b_1^{**})$ (because it lies in $\mathbf{B}_1$). In particular, this means $b_1^{**'} \leq b_1^{**}$, and more agents share on island 1 in the most-sharing equilibrium following the rise in homophily.

To measure the change in virality, we first observe that the seed agent i^* (that maximizes $\mathbb{E}[\mathbf{S}_{i^*}]$) is chosen from the agents on island 1. We consider the virality of the article when agents on island 1 share with probability $1 - H(b_1^{**})$ under the stronger homophily structure (p_s', p_d') versus (p_s, p_d) (and all other agents kill the article). This is sufficient to show that virality increases following the increase in homophily, because virality with $b_1^{**'} < b_1^{**}$ (but the same network $\mathbf{P}$) is strictly higher, given that agents on island 1 share more often, that is, $(1 - H(b^{**'}) > 1 - H(b^{**}))$.

We consider the diffusion process of an article on the (p_s', p_d') network that starts with an agent on island 1. Let us define a *path* of the diffusion process to be a chain $i^* \to i_1 \to i_2 \to \ldots \to i_z$ representing a sequence of agents who receive the article in this process, with i^* being the seed agent, i_1 through i_{z-1} all being agents who shared it, and agent i_z being an agent who either ignored or disliked the article. There may be many such paths for the diffusion of the article (by assumption, all agents with the possible exception of agent i_z must be on island 1).

For each path, we define an alternative path (generated randomly) as follows. For any links to agents other than to agent i_z (i.e., links within island 1), with probability $(p_s' - p_s)/p_s'$, the link instead goes to one of islands $2, \ldots, k$ (chosen in proportion to their population) and otherwise remains the same. Applying this to all paths, we define an isomorphic diffusion process to one on a sharing network with weaker homophily parameters (p_s, p_d). However, note that the length of every path cannot increase following this transformation. Because any transition to islands $2, \ldots, k$ is necessarily the end of the path, paths can only shorten. Moreover, the number of paths must weakly decrease. As a result, the fraction of agents who receive the article, $\mathbf{S}_{i^*}$, must be lower, and virality is less under the (p_s, p_d) sharing network. This establishes part (a).

For part (b), we first note that there exists $\bar{r}$ such that the most-sharing equilibrium when $r > \bar{r}$ is all-share ($b_\ell^{**} = 0$ for all islands ℓ) regardless of $\mathbf{P}$. Notice that equation (2.2) is minimized when $b_i = 0$, and in particular, for all agents i (regardless of their prior) $\pi_i \geq \frac{(1-p)\phi(r)}{(1-q)(1-\phi(r))+(1-p)\phi(r)}$.

Then, letting $\bar{\pi} = \max\left\{\frac{c}{u+c}, \frac{\tilde{u}-\tilde{c}}{\tilde{u}}\right\} < 1$, we note that whenever $r \geq \phi^{-1}\left(\frac{(1-q)\bar{\pi}}{(p-q)\bar{\pi}+(1-p)}\right) \equiv \bar{r} \in (0,1)$, $\pi_i \geq \bar{\pi}$. Of course, when all other agents (other than i) share and $r > \bar{r}$, $U_i(\mathcal{S}) \geq u\pi_i - c(1-\pi_i) \geq 0$ and $U_i(\mathcal{D}) = \tilde{u}(1-\pi_i) - \tilde{c} \leq 0$, so $a_i = \mathcal{S}$ is a best response for agent i. Thus, the most-sharing equilibrium is all-share (because it *is* an equilibrium and no other strategy profile can have more sharing).

Observe that when $r > \bar{r}$, virality is measured simply by the expected size of the connected component (formed by $\mathbf{P}$) containing the seed agent i^*. Regardless of the homophily parameters, the seed agent i^* will be chosen from the largest island (call this island ℓ^*). This is immediate from the fact that all agents share in equilibrium, agents on island ℓ^* have the most connections to any other arbitrary island ℓ' (in expectation), and are connected to all agents on their own island.

Lastly, we note that the probability that island ℓ has any connections to island ℓ' is given by $\tilde{p}_{\ell,\ell'} = 1-(1-p_d)^{N_\ell N'_\ell}$ *before* the decrease in homophily and $\tilde{p}'_{\ell,\ell'} = 1-(1-p'_d)^{N_\ell N'_\ell}$ *after* the decrease in homophily, with $\tilde{p}'_{\ell,\ell'} > \tilde{p}_{\ell,\ell'}$ for all pairs of islands (ℓ,ℓ') because $p'_d > p_d$. Using the same terminology as in the argument for part (a), we map the diffusion paths of an article under the less homophilic sharing network with (p'_s, p'_d). Consider cycles between islands $\ell^* \to \ell_1 \to \ell_2 \ldots \to \ell_z$, where ℓ_z is the same island as one of $\ell^*, \ell_1, \ldots, \ell_z$ (in which case, no additional engagement is obtained thereafter the article returns to island ℓ_z). Before the decrease in homophily (where $p_d < p'_d$), we can construct an isomorphic diffusion process where an article remains within the same island (instead of switching to a different one) with probability $(p'_d - p_d)/p_d$. By construction of the cycle, whenever such an event occurs, the cycle becomes complete and the islands reached thereafter in the (p'_s, p'_d) sharing network are not (for that given cycle). Measuring across all cycles that occur in the (p'_s, p'_d) model, (weakly) more islands are reached than under the more homophilic (p_s, p_d) model. Consequently, virality is higher under the (p'_s, p'_d) sharing network than with the (p_s, p_d) sharing network, which has more homophily. This establishes part (b). ∎

Proof of Proposition 2.1.2. Let us define r^* as

$$r^* \equiv \phi^{-1}\left(\max\left\{\frac{(1-q)(\tilde{u}-\tilde{c})}{(p-q)(\tilde{u}-\tilde{c})+(1-p)\tilde{u}}, \frac{c}{u+c}\right\}\right) \in (0,1)\,.$$

For part (a), first consider the case of $r < r^*$ and $p_d = 0$ (by continuity, the result extends to the case of sufficiently large p_s/p_d). In the most-sharing equilibrium, the seed agent most conducive to the article's spread is on the right-wing island, and given that $p_d = 0$, the equilibrium on the left-wing island is immaterial to its virality. Let us denote the right-wing island cutoffs by (b^*_R, b^{**}_R). Similar to the proof of Theorem 2.1.2(a), we define a cutoff space $\mathbf{B}_R$ such that $(\hat{b}^*_R, \hat{b}^{**}_R) \in \mathbf{B}_R$ if and only if $(\hat{b}^*_R, \hat{b}^{**}_R) \preceq (b^*_R, b^{**}_R)$. Similarly, we define the map $\varphi : \mathbf{B}_R \to \mathbf{B}_R$ which maps an arbitrary cutoff $(\hat{b}^*_R, \hat{b}^{**}_R)$ to best-response cutoffs $(\hat{b}^{*,BR}_R, \hat{b}^{**,BR}_R)$. To show the map is well-defined, consider $U_R(\mathcal{S})$ *before* the increase in divisiveness or polarization and $U'_R(\mathcal{S})$ *after* the increase in divisiveness or polarization. Because the network structure is fixed, note that $U^{(2)}_R(\mathcal{S}) = U^{(2)'}_R(\mathcal{S})$ when the cutoffs $(\hat{b}^*_R, \hat{b}^{**}_R)$ are taken as given, so the difference $U'_R(\mathcal{S}) - U_R(\mathcal{S})$ depends only on the difference between $U^{(1)}_R(\mathcal{S})$ and $U^{(1)'}_R(\mathcal{S})$. Specifically, the difference in share payoff depends only on the change in π_i following the increase in divisiveness or polarization. Moreover,

$$\frac{\partial \pi_i}{\partial p} = \frac{(2b_i - 1)(1 - \phi(r))\phi(r)(1 - q - b_i(1 - 2q))}{(b_i(2p\phi(r) + 2q(1 - \phi(r)) - 1) - p\phi(r) - q(1 - \phi(r)) + 1)^2} > 0\,;$$
$$\frac{\partial \pi_i}{\partial q} = \frac{(2b_i - 1)(1 - \phi(r))\phi(r)(1 - p + b_i(2p - 1))}{((2b_i - 1)\phi(r)(p - q) + 2b_i q - b_i - q + 1)^2} > 0\,,$$

whenever $b_i > 1/2$. Likewise, as we showed in Lemma A.1.1, $\partial\pi_i/\partial b_i > 0$ for all b_i and greater polarization increases ideological priors for agents with $b_i > 1/2$ (by Lemma A.1.3). By virtue of $\underline{b}_R > 1/2$, we observe that $U^{(1)'}_R(\mathcal{S}) > U^{(1)}_R(\mathcal{S})$, and so $U'_R(\mathcal{S}) > U_R(\mathcal{S})$. Thus, as in the proof of Theorem 2.1.2(a), φ is well-defined. Applying the Tarski fixed-point theorem, we find that the most-sharing equilibrium leads to more sharing in the right-wing island. Because the network structure $\mathbf{P}$ remains constant and there is a uniform shift in sharing, our weaker notion of virality also increases.

For part (b), consider $r \geq r^*$. Note that for $r \geq r^*$, ignoring is a better response to disliking for any agent, regardless of prior and sharing is a better response to ignoring for all $b_i > 1/2$. The former follows from noting $\pi_i \geq \frac{\tilde{u}-\tilde{c}}{\tilde{u}}$ for an agent with prior $b_i = 0$ and the latter from noting $\pi_i \geq \frac{c}{u+c}$ for agents with $b_i > 1/2$ *and* observing that disliking is a dominated strategy. Therefore, the right-wing island always shares, whereas the left-wing island has equilibrium cutoffs $(0, b^{**}_L)$. Using the same approach as in part (a), it is enough to show that $U_L(\mathcal{S})$ increases following a decrease in divisiveness or polarization. Furthermore, for $b_i < 1/2$, we see that $\partial\pi_i/\partial p < 0$ and $\partial\pi_i/\partial q < 0$, and by Lemma A.1.3, decreasing polarization means that all agents on the left-wing island also have an increase in b_i. Thus, there is more sharing in the most-sharing equilibrium following a decrease in divisiveness or polarization. By strategic complementarity, the right-wing island remains at all-share, and sharing uniformly increases (and so does virality, naturally). ▮

Proofs from Section 5

Proof of Theorem 2.1.3. Consider the complete sharing network where $\mathbf{P} = \mathbf{1}_{N\times N} - \mathbf{I}$. We claim that if the most-sharing equilibrium involves all agents choosing $\mathcal{S}$ or $\mathcal{I}$ (with probability 1) and agents never choosing $\mathcal{D}$ under this configuration, then this is the platform's profit-maximizing sharing network. By Lemma 2.1.1, all agents employ the same cutoffs $(b^*, b^{**}) = (0, b^{**})$ and $1 - H(b^{**})$ determines the proportion who share in the most-sharing equilibrium.

We focus on a modified social media game that only allows agents to ignore or share, which necessarily increases virality of content for any sharing network $\mathbf{P}'$ but does not increase the virality for the complete sharing network, by assumption. We show that for any other sharing network $\mathbf{P}'$, the largest fixed point (the most-sharing equilibrium), must necessarily be above $b^{**}\mathbf{1}$ (in the order $\preceq$). To do this, we consider the largest fixed point under $\mathbf{P}'$ (call this $\mathbf{b}^{**'}$), and use the same mathematical arguments as before, only disregarding the dislike cutoff. Let $\mathbf{B}'$ be the cutoff space where $\hat{\mathbf{b}}^{**}$ satisfies $\hat{\mathbf{b}}^{**} \preceq \mathbf{b}^{**'}$ and let the map $\varphi : \mathbf{B}' \to \mathbf{B}'$ map fixed cutoff strategies $\hat{\mathbf{b}}^{**}$ to best-response sharing cutoff strategies under the complete sharing network. It only remains to prove that φ indeed maps into $\mathbf{B}'$. To do this, let $U_j^c(\mathcal{S})$ be the utility from sharing under the complete network and $U_j'(\mathcal{S})$ as sharing under $\mathbf{P}'$, and note that $U_j^c(\mathcal{S}) - U_j'(\mathcal{S}) = \kappa \sum_{\tilde{j}=1}^{N} (1 - p'_{j\tilde{j}})(1 - H(\hat{b}^{**}_{\tilde{j}})) \geq 0$.

Thus, by Tarski's fixed-point theorem, we once again obtain that $b^{**}\mathbf{1} \preceq \mathbf{b}^{**'}$. Finally, observe that this necessarily implies that $\mathbf{P}'$ is less viral, because for every prior realization and seed agent i^*, $\mathbf{S}_{i^*}$ is larger in the complete network than in any other sharing network, provided that $b^{**}\mathbf{1} \preceq \mathbf{b}^{**'}$ and $b^* = 0$ (no agent dislikes). By Proposition 2.1.1, (uniformly more) sharing is monotone in reliability, so there exists some r_P such that for $r > r_P$, the complete sharing network admits only shares and ignores, whereas when $r < r_P$, agents dislike with positive probability. When $r > r_P$, the network takes the form of part (ii) by setting $p_s = p_d = 1$.

Next, we consider the case where $r < r_P$, and so (b^*, b^{**}) are the cutoffs in the most-sharing equilibrium with a complete sharing network, but where $b^* > 0$. First, notice that there must exist an open interval where agents with priors $b_i \in (0, \bar{b})$ will *never* share, regardless of the sharing network $\mathbf{P}$. To see this, suppose that there exists some $\mathbf{P}'$ where *all* agents either share or ignore, so the equilibrium cutoffs are determined by $\mathbf{b}^{**'}$ (and $\mathbf{b}^{*'} = \mathbf{0}$). Using the reasoning as in the previous paragraph (but extending it to the full cutoff space $(b^*\mathbf{1}, b^{**}\mathbf{1})$), we conclude that the cutoffs in the most-sharing equilibrium of the complete sharing network must satisfy $(b^*\mathbf{1}, b^{**}\mathbf{1}) \preceq (\mathbf{b}^{*'}, \mathbf{b}^{**'})$, and in particular, $b^*\mathbf{1} \preceq \mathbf{b}^{*'} = \mathbf{0}$. This implies that all agents share or ignore in the complete network, yielding a contradiction. Therefore, such an interval $(0, \bar{b})$ must exist, and in particular, we choose the largest such $\bar{b}$ (in the supremum sense) where agents with priors in $(0, \bar{b})$ *never* share in any sharing network $\mathbf{P}$ in the most-sharing equilibrium.

Next, we consider disconnecting (and removing) all agents in any community ℓ with $b^{(\ell)} < \bar{b}$, but leaving all other communities connected in a (partial) complete network. We call this network the *active network.* We claim that when ε is sufficiently small, all of the remaining agents in the active network either share or ignore. By definition, an agent with $\bar{b}$ would share under some sharing network $\mathbf{P}^*$ but any agent with $\bar{b} - \epsilon$ would ignore (for arbitrarily small ϵ) under $\mathbf{P}^*$ (and by leveraging Lemma A.1.4(i), not all agents with $b < \bar{b}$ dislike). This implies that $U_i(\mathcal{D}) < 0$ for an agent with prior $\bar{b}$ (by monotonicity), and thus an agent with this prior either shares or ignores in the active network. Moreover, for agents with priors in a small half-open neighborhood around $\bar{b}$ (i.e., an interval $(\bar{b} - \eta, \bar{b}]$ for some $\eta > 0$) ignoring is a better response to disliking in the active network. Thus, for sufficiently small ε, we obtain a partial complete network (the active network) with agents who only share and ignore (with probability 1) and never dislike (with probability 0).

Finally, with these two observations, we claim that the profit-maximizing sharing network takes the form of part (i). First, consider all communities who participated in the active network described above (call these the active communities) and suppose that the agents in communities outside of this active network are non-existent in our model (call these the inactive communities). When the active communities are arranged in a complete sharing network, we showed in the previous paragraph that all agents either share or ignore. By the exact argument in the first two paragraphs then, engagement (and virality) are maximized (amongst only the active communities) when these communities are arranged in a complete sharing network. Second, by construction of the active network (and the active communities), all agents in inactive communities *never* share under *any* sharing network $\mathbf{P}$. Therefore, removing these agents is without loss to potential virality. Hence, whenever virality is maximized amongst only agents in active communities, it is also maximized in general.

Lastly, we note that we can form a (partial) complete network among the inactive communities, but provide no connections to the (partial) complete network of active communities who only share and ignore (with probability 1). By our previous observations, this is a profit-maximizing sharing network for the platform. At the same time, it is exactly the form of a two-island model with $(p_s, p_d) = (1, 0)$, which has maximal homophily. ∎

Proof of Proposition 2.1.3. Note by Theorem 2.1.3 that an agent i with prior $b_i = b^{(k+1)}$ is indifferent between ignoring and disliking when $r = r_P$ (but strictly prefers to either share or ignore for all $r > r_P$), so r_P increases if and only if this agent (strictly) prefers to dislike following a shift in parameters. Because $b^{(k+1)} < 1/2$, an increase in polarization means that agent i's prior decreases (see Lemma A.1.3), and given that $\partial \pi_i / \partial b_i > 0$ (see Lemma A.1.1), π_i decreases for this agent. As a consequence $U_i(\mathcal{D})$ increases but $U_i(\mathcal{I})$ remains the same, so agent i (strictly) prefers to dislike. Similarly, because $\partial \pi_i / \partial p < 0$ and $\partial \pi_i / \partial q < 0$ for $b_i < 1/2$ (see Proposition 2.1.2), π_i decreases for this agent (making $a_i = \mathcal{D}$ a best response). In both cases, we see that r_P increases. ∎

Proofs from Section 6

Proof of Proposition 2.1.4. Consider the profit-maximizing sharing network before any censorship policy is enacted ($\delta = 0$). By Theorem 2.1.3 and the assumption that $\mathbf{b}^* \neq \mathbf{0}$ and $\mathbf{b}^{**} \neq \mathbf{1}$, it must be the case the profit-maximizing sharing network has maximal homophily with two islands, one with the optimal seed agent (island A) and one without it (island B). By construction of the profit-maximizing sharing network, no agent on island B would share if connected fully to island A or under any other sharing network configuration (see the proof of Theorem 2.1.3).

Consider any agent j residing on island B. Because the platform approximates the belief distribution H by a generic multinomial distribution,[2] it must be the case that $U_j(\mathcal{S}) < \max\{U_j(\mathcal{I}), U_j(\mathcal{D})\}$ under any sharing network configuration for agent j. Hence, there exists some $\underline{\delta} > 0$ such that substituting $\phi(r)$ with $\tilde{\phi}(r) = \frac{\phi(r)}{\phi(r)+(1-\delta)(1-\phi(r))}$ for all $\delta \in (0, \underline{\delta})$ leaves the

[2] Formally, given that the platform has microtargeting technology $\varepsilon > 0$, the optimally chosen sharing network (for the platform) with prior distribution H is equivalent to the platform's optimally chosen sharing network for a multinomial distribution consisting of a number of atoms chosen "generically" (each atom chosen at random from an interval of size ε) in each of the prior regions $[b^{(k)}, b^{(k+1)}]$, as described in Section 2.1.5.

profit-maximizing sharing network for the platform unchanged, because the strict inequality above still holds under any chosen sharing network. At the same time, if $E \equiv \max_{i^*} \mathbb{E}[\mathbf{S}_{i^*}|\nu = \mathcal{M}]$ is the engagement with misinformation before the policy, engagement with misinformation after the policy is $(1-\delta)E < E$. Thus, the policy for $\delta^* \in (0, \underline{\delta})$ is more effective than $\delta = 0$, and in fact higher values of $\delta^* \in (0, \underline{\delta})$ are more effective.

Next, we note that $\tilde{\phi}(r) = \frac{\phi(r)}{\phi(r)+(1-\delta)(1-\phi(r))} = 1$ when $\delta = 1$. It is immediate then that for sufficiently high values of δ, the profit-maximizing sharing network has maximal connectivity and all agents share in equilibrium. Let us consider $\bar{\delta}$ which is the largest value (in the supremum sense) such that the profit-maximizing sharing network does not have maximal connectivity. Under censorship policy $\bar{\delta}$, the virality of misinformation is at most $(1-\bar{\delta})(N-1)/N$, but for any $\zeta > 0$, the censorship policy with $\bar{\delta}$ has virality $1 - \bar{\delta} - \zeta$. Letting $\zeta < (1-\bar{\delta})/N$, we see that a censorship policy with $\bar{\delta}$ is more effective than any policy with $\bar{\delta} + \zeta$.

Finally, let us construct $0 < \delta_1 < \delta_2 < \delta_3 < 1$ to conclude. Let us take δ_1 to be the largest value of δ (in the supremum sense) such that $\delta^* = \delta$ is the most effective policy for all $\delta \in (0, \delta_1)$. We know that that such a δ_1 exists and is strictly less than 1 by the arguments in the two above paragraphs. Using a similar argument as in the second paragraph, there always exists an open interval (δ_1, δ_2) where the $\delta^* = \delta_1$ policy is more effective than any $\delta^* \in (\delta_1, \delta_2)$, and so in particular $\delta^* < \delta$ is optimal. Lastly, we show (i) any censorship policy bounded away from 1 can always be beat by one sufficiently close to 1, and (ii) the one that beats it can beat by any δ greater than it. This proves that there exists δ_3 whenever $\delta \in (\delta_3, 1)$, $\delta^* = \delta$ is the most effective policy. For (i) suppose that some policy $\tilde{\delta}$ achieves misinformation engagement $\tilde{E} > 0$; then, because engagement with unidentified misinformation cannot exceed N, any policy $\delta^* > (N - \tilde{E})/N$ is strictly more effective. For (ii) we know there exists some $\hat{\delta}$ sufficiently close to 1 such that profit-maximizing sharing network has maximal connectivity and thus is the same for all $\delta^* \in (\hat{\delta}, 1)$; therefore, the virality of misinformation is $(1-\delta^*)$ for all $\delta^* \in (\hat{\delta}, 1)$ and higher values of δ^* are always more effective. ∎

Proof of Proposition 2.1.5. We take a similar approach as in the proof of Proposition 2.1.4. Once again, we consider islands A and B which are guaranteed by Theorem 2.1.3 before any provenance policy has been enacted ($\rho = 0$) and note that $U_j(\mathcal{S}) < \max\{U_j(\mathcal{I}), U_j(\mathcal{D})\}$ for all agents j on island B regardless of the sharing network chosen. With the introduction of a provenance policy, however, the profit-maximizing sharing network may not take the form of

Theorem 2.1.3. Despite this, we can still upper bound the ex-ante likelihood of an article being truthful by $\frac{\phi(r)}{\phi(r)+(1-\rho)^N(1-\phi(r))}$, which holds independent of the sharing network chosen. Once again, for small enough $\underline{\rho} > 0$ the strict inequality $U_j(\mathcal{S}) < \max\{U_j(\mathcal{I}), U_j(\mathcal{D})\}$ will still hold in any sharing network, and the profit-maximizing sharing network will be the same as $\rho = 0$ for all $\rho^* \in (0, \underline{\rho})$. In this region, virality is given by $[(1-\rho^*) + (1-\rho^*)^2]\,\beta$ where β is the fraction of agents (relative to N) on island A, which is monotonically decreasing in ρ^*.

At the same time, we can lower bound the ex-ante likelihood of an article being truthful by $\frac{\phi(r)}{\phi(r)+(1-\rho)(1-\phi(r))}$, which is equal to 1 when $\rho = 1$. Thus, note that for sufficiently high values of ρ, the profit-maximizing sharing network will fit the form of Theorem 2.1.3 with a maximally connected network because all agents share in equilibrium achieving maximal virality. Once again, let us consider $\bar{\rho}$ which is the largest value (in the supremum sense) such that the profit-maximizing sharing network is something *other* than maximal connectivity. For $\rho = \bar{\rho}$, there must be at least one agent who would not share under any chosen sharing network. As in Proposition 2.1.4, for any $\zeta < (1-\bar{\rho})/N$, a provenance policy with $\bar{\rho}$ is more effective than any policy with $\bar{\rho} + \zeta$. The construction of $0 < \rho_1 < \rho_2 < \rho_3 < 1$ then follows exactly in the same way as from the last paragraph in Proposition 2.1.4.

Finally, we show that $\rho_3 \leq \delta_3$ and in this region the provenance policy is more effective than censorship. Recall that δ_3 was chosen such that it is the minimum value of δ where the profit-maximizing sharing network is maximally connected, and for $\rho = \delta_3$, it must also be maximally connected, because the perceived ex-ante likelihood of truth is lower bounded by $\frac{\phi(r)}{\phi(r)+(1-\rho)(1-\phi(r))}$, which is the ex-ante likelihood of truth for a censorship policy where $\delta = \rho$. The expected virality in the censorship regime for $\delta^* > \delta_3$ is $(1-\delta^*)$, but it is only $[(1-\rho^*) + (1-\rho^*)^2(N-1)]/N < (1-\rho^*)$ when $\rho^* > \rho_3$. ∎

Proof of Proposition 2.1.6. If the platform removes ψ fraction of misinformation, then its performance metric is given by

$$\frac{(1-\psi)E(\psi)(1-\phi(r))}{(1-\psi)E(\psi)(1-\phi(r)) + E(\psi)\phi(r)} = \frac{(1-\psi)(1-\phi(r))}{(1-\psi)(1-\phi(r)) + \phi(r)}$$

where $E(\psi)$ is the user engagement when the platform removes ψ fraction of misinformation and optimally chooses the sharing network, but notice that $E(\psi)$ does not affect the performance metric.

If the platform hits the performance target λ, then it chooses ψ according to $\psi = \frac{1-\lambda-\phi(r)}{(1-\lambda)(1-\phi(r))}$. Observe that ψ is monotonically decreasing in λ, with the strictest target ($\lambda = 0$) yielding $\psi = 1$ and the loosest target ($\lambda = 1 - \phi(r)$) yielding $\psi = 0$. The payoff from hitting the performance target exactly is given by $V = E(\psi)\phi(r)/(1-\lambda)$ whereas the payoff from not hitting it is $V' = (1-\alpha)E(\psi^*) - \alpha C$, where ψ^* is the self-imposed target by the platform that maximizes engagement, i.e., $\psi^* = \max_\psi E(\psi) < 1$ by nature of $\phi(r) < \max_{i^*} \mathbb{E}[\mathbf{S}_{i^*}]$. For any $\psi < \psi^*$, the platform of course meets the performance target. For any $\psi > \psi^*$, we can define $E^*(\psi) = \max_{\psi' \geq \psi} E(\psi)$, which is of course a monotonically decreasing function in ψ. Thus, the platform compares $V = E^*(\psi)\phi(r)/(1-\lambda)$ with a constant $V' = (1-\alpha)E(\psi^*) - \alpha C$. Note that V is monotonically increasing in λ: $1/(1-\lambda)$ is increasing in λ, and $E^*(\psi)$ is decreasing in ψ, and therefore increasing in λ. Thus, there exists some cutoff λ^* such that when $\lambda > \lambda^*$, $V > V'$, but when $\lambda < \lambda^*$, $V < V'$. The claim follows by noting that virality of misinformation is proportional to $E(\psi)$ where ψ is chosen by the platform. ∎

Proof of Proposition 2.1.7. The network regulation does not bind for an article with $r > r_P$, so we need only consider $r < r_P$. Take some agent i with prior $b_i \in (\bar{b}, \bar{b} + \eta)$ in a small neighborhood $\eta > 0$ of $\bar{b}$ (where $\bar{b}$ is the same $\bar{b}$ constructed in Theorem 2.1.3). Following the same line of reasoning as in Theorem 2.1.2(a), agents with priors in this interval elect to ignore instead of share following the network regulation (and when η is sufficiently small), and this necessarily reduces the virality of misinformation, showing (ii). To prove (i), we note that agents in this neighborhood around $\bar{b}$ also do not share in the most-sharing equilibrium under *any* sharing network $\mathbf{P}'$ (following the network regulation), per the construction of $\bar{b}$ in Theorem 2.1.3. Therefore, the platform cannot generate additional engagement by departing from the class of island models (specifically, two-island models) while maintaining $p_s/p_d \leq p^*$. ∎

A.1.2 Endogenous Reputation Loss

In the model of Section 5.2, we assume that each agent cares about getting called out for sharing potential misinformation, in the form of exogenous punishments for each dislike she receives. In this section, we show that this formulation can be microfounded by an endogenous reputational concern.

Suppose that at $t = 0$, every agent is born as either a *careless* agent or a *normal* agent. The careless type is behavioral and shares every article, whereas a normal user is a fully rational agent as in Section 5.2. The user is born careless with probability $\mu > 0$ which is i.i.d. across all agents and immutable. For each dislike agent i receives, there is some probability $\zeta > 0$ that the share by agent i receives public scrutiny and it becomes common knowledge to the population that i shared the article. Conditional on such a broadcast, the population updates its beliefs μ to $\hat{\mu}_i$ about the likelihood that agent i is actually the careless type.

We assume each agent i intrinsically values $1 - \hat{\mu}_i$, the public belief that she is not a careless user, for example, because this might affect their other social relations or economic prospects. We can represent the strength of this concern (relative to other sources of utility) with a parameter like d in Section 5.2. For example, a doctor may place much more weight on reputation than a social media troll trying to "stir the pot."

When there is no public broadcast about agent i, we have $\hat{\mu}_i \leq \mu_i$. In contrast, when there is a broadcast about i's share, $\hat{\mu}_i > \mu_i$, and this increase is larger for articles with lower reliability. This reasoning therefore introduces an endogenous reputational concern originating from dislikes that the agent receives. Specifically, it is straightforward to see that the agent's utility will now include a term $\psi_D(d, D_i) = d\Delta\hat{\mu}_i(1 - (1 - \zeta)^{D_i})$, where $\Delta\hat{\mu}_i$ is the difference between i's reputation after a public broadcast of her share and i's reputation without a broadcast, and $\psi_D(d, D_i)$ exhibits increasing differences. By the same observation as in footnote 7, all of our results apply without modification.

A.2 Demand for Misinformation: Fighting Fire with Fire

A.2.1 Proofs

Appendix B

Misinformation: Behavioral Models

B.1 DeGroot Models I: Manipulating with Misinformation

B.1.1 Technical Details

Local Density. We provide a generalization of Theorem 3.1.4, which is often more useful in practice, especially when stubborn agents "disconnect" the network (i.e., there exist DeGroots i, j with the only directed walks between them containing stubborn agents). In Example B.1.3 (see Appendix B.3.2), we apply the result to one variant of the star network.

Definition B.1.1. The log-distance between i and j is:

$$d_{ij} = \min_{W_{ij} \in \mathcal{W}_{ij}} \sum_{(i' \to j') \in W_{ij}} -\log(w_{i'j'})$$

We say that network G is δ-locally dense if there exist subsets $I_1, \ldots, I_k$ of agents in G such that: (i) $\cup_{\ell=1}^{k} I_\ell = \{1, \ldots, n\}$ (i.e., the subsets cover G) and (ii) the log-distance between every two agents $i, j \in I_\ell$ is at most $\log(|I_\ell| + \delta)$.

Proposition B.1.1. *If the network* G *is* δ*-locally dense and contains* $m^*(\delta)$ *stubborn agents (from Theorem 3.1.4) in each set* I_ℓ*, the network is impervious.*

It is easy to see Theorem 3.1.4 is a special case of Proposition B.1.1 by taking $I_1 = \{1, \ldots, n\}$ and checking the log-distance between every two agents in G (i.e., log-diameter) is at most $\log(|I_1| + \delta) = \log(n + \delta)$.

Full Characterization of Principal's Problem. We can write the principal's problem as the following integer (binary) program:

$$\begin{aligned} \mathbf{\Gamma}^* &= \arg\max \sum_{i=m+1}^{n} r_i - \varepsilon\gamma_i \\ \text{s.t. } \forall i: \; & r_i \leq \mathcal{D}_i(\boldsymbol{\gamma}) + (1+b)/2 \\ \forall i: \; & \gamma_i, r_i \in \{0,1\} \end{aligned}$$

Theorem B.1.1. *Given investment cost $\varepsilon > 0$ and a solution $\mathbf{\Gamma}^*$ to the principal's problem, a network is impervious if $\mathbf{0} \in \mathbf{\Gamma}^*$; otherwise it is susceptible.*

The principal can choose to either send misinformation ($\gamma_i = 1$) or not ($\gamma_i = 0$) for each agent. The choice of $\boldsymbol{\gamma}$ impacts the principal's payoffs in two ways: (i) a direct, separable cost ε for each $\gamma_i = 1$ and (ii) a network impact captured in the DeGroot centrality (i.e., how the experiences of DeGroot agents impact the beliefs of others) from the aggregate vector $\boldsymbol{\gamma}$. In Appendix B.1.4, we use this problem to solve explicitly for the optimal strategy in a real-world social network.

Note that $\mathcal{D}_i(\boldsymbol{\gamma})$ is *linear* in $\boldsymbol{\gamma}$, which makes the problem an integer program (IP) for any network G. Despite this, such an optimization problem is generally intractable. However, we can provide sufficient conditions for showing that a network is either impervious or susceptible to manipulation. These conditions, for most networks in practice, tend to be much more useful than direct application of this optimization problem. For notation purposes, for a subset $\mathcal{K} \subset D$ of DeGroot agents let $\mathbf{1}_{\mathcal{K}}$ denote the vector given by:

$$[\mathbf{1}_{\mathcal{K}}]_i = \begin{cases} 1, & \text{if } i \in \mathcal{K} \\ 0, & \text{otherwise} \end{cases}$$

Then we obtain the following corollary to Theorem B.1.1:

Corollary B.1.1. *Fix some $\varepsilon > 0$; then the network is:*

(a) Impervious *to manipulation if $\mathcal{D}_i(\mathbf{1}_D) < (1-b)/2$ for every DeGroot agent i, or*

(b) Susceptible *to manipulation if there exists a subset $\mathcal{K} \neq \varnothing$ of DeGroot agents such that:*

$$\sum_{i=m+1}^{n} 1_{\mathcal{D}_i(\mathbf{1}_{\mathcal{K}})>(1-b)/2} > \varepsilon|\mathcal{K}|$$

Note that the condition on imperviousness is sufficient but not necessary. It simply states that if the principal sends signals to all of the DeGroot agents, the influence from the stubborn agents will still dominate (i.e., ensure DeGroots take the correct action). We see this result holds regardless of the cost of investment ε; in particular, it becomes a necessary condition as well when $\varepsilon \to 0$. However, a necessary *and* sufficient condition for susceptibility is given by (b). While it is challenging to verify that there exists no subset $\mathcal{K}$ that is profitable for the principal to manipulate, it is often easy to simply check that some subset $\mathcal{K}$ does better than $\gamma = \mathbf{0}$.

B.1.2 Proofs

Section 3

Proof of Lemma 3.1.1. We first prove that $\text{BU}(S|h_{i,t})$ is a martingale. Consider the filtration with respect to the history $h_{i,t}$. Then:

$$\begin{aligned}\mathbb{E}\left[\text{BU}(S|h_{i,t+1})|h_{i,t}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\theta=S}|h_{i,t+1}\right]|h_{i,t}\right]\\ &= \mathbb{E}[\mathbf{1}_{\theta=S}|h_{i,t}]\\ &= \text{BU}(S|h_{i,t})\end{aligned}$$

where the second to last inequality follows from the law of iterated expectations. Because the Bayesian update term is a belief and bounded between 0 and 1, we know by the martingale convergence theorem that $\text{BU}(S|h_{i,t})$ converges almost surely to a random variable X. We next prove that X is a constant almost surely. If $p_i = 1/2$, then $\text{BU}(S|h_{i,t}) = \text{BU}(S|h_{i,0}) = q$ for all t and so trivially converges to constant q. Otherwise, we know that if $\gamma_i = 1$ then DeGroot agent i receives signal R with probability $\frac{\lambda^*}{\lambda+\lambda^*} + \frac{\lambda}{\lambda+\lambda^*}(1-p_i) > 1/2$ by Assumption 3.1.1. We show that $\text{BU}(S|h_{i,t})$ converges almost surely to 0. Consider the biased random walk $z^{\Delta}_{i,t} = z^{R}_{i,t} - z^{S}_{i,t}$. For all t we can write:

$$\begin{aligned}
\mathrm{BU}(S|h_{i,t}) &= \frac{p_i^{z_{i,t}^S}(1-p_i)^{z_{i,t}^R} q}{p_i^{z_{i,t}^S}(1-p_i)^{z_{i,t}^R} q + p_i^{z_{i,t}^R}(1-p_i)^{z_{i,t}^S}(1-q)} \\
&= \frac{q}{q + \left(\frac{p_i}{1-p_i}\right)^{z_{i,t}^{\Delta}}(1-q)} \\
&\overset{a.s.}{\to} 0
\end{aligned}$$

because for a biased random walk with the probability of R greater than $1/2$, we know that $z_{i,t}^{\Delta} \overset{a.s.}{\to} \infty$, and $p_i > 1/2$.

Similarly, if $\gamma_i = 0$, then DeGroot agent i receives signal S with probability $p_i > 1/2$. We show that $\mathrm{BU}(S|h_{i,t})$ converges almost surely to 1. Consider the same biased random walk; then for all t we can write:

$$\begin{aligned}
\mathrm{BU}(S|h_{i,t}) &= \frac{p_i^{z_{i,t}^S}(1-p_i)^{z_{i,t}^R} q}{p_i^{z_{i,t}^S}(1-p_i)^{z_{i,t}^R} q + p_i^{z_{i,t}^R}(1-p_i)^{z_{i,t}^S}(1-q)} \\
&= \frac{q}{q + \left(\frac{1-p_i}{p_i}\right)^{-z_{i,t}^{\Delta}}(1-q)} \\
&\overset{a.s.}{\to} 1
\end{aligned}$$

because for a biased random walk with the probability of S greater than $1/2$, we know that $-z_{i,t}^{\Delta} \overset{a.s.}{\to} \infty$, and $p_i > 1/2$. ∎

Lemma B.1.1. *The spectral radius of matrix* $\mathbf{W}$ *is strictly less than 1.*

Proof. It is equivalent to prove that all the eigenvalues of $\mathbf{W}$ lie strictly within the unit circle. For stubborn agents or DeGroot agents with $\theta_i = 1$, these agents have $\alpha_{ij} = 0$ for all j, so $\mathbf{W}_i$ is the zero vector. Thus, these agents introduce an additional eigenvalue of 0, which of course lies within the unit circle, without affecting the rest of the eigenvalues. Therefore, it is without loss of generality to consider DeGroot agents with $\theta_i < 1$ for all agents i (where we assign arbitrary θ values for those with θ equal to 1 (as we have already identified this as irrelevant.) Then let us define the diagonal matrix:

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{1-\theta_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{1-\theta_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{1-\theta_n} \end{pmatrix}$$

Then we note that $\mathbf{QW}$ is row-stochastic, so by the Perron-Frobenius theorem all eigenvalues lie strictly within the unit circle except for the largest, which is exactly equal to 1. Further, because there is at least one agent with $p_i > 1/2$, by Assumption 3.1.1, this agent is either stubborn or DeGroot with $\theta_i > 0$, so $\mathbf{Q}$ has at least one eigenvalue strictly greater than 1, with corresponding eigenvector $\mathbf{v}^*$. Moreover, none of the eigenvalues of $\mathbf{Q}$ are less than or equal to 1.

Consider any arbitrary vector $\mathbf{v} \in \mathbb{R}^{n-m}$. By Assumption 3.1.1 (strong connectedness), we know there exists k such that $\mathbf{QW}^k\mathbf{v}$ is not a scalar-multiple of $\mathbf{v}^*$, and so we obtain the strong inequality:

$$||\mathbf{W}^k\mathbf{v}||_2 < ||\mathbf{QW}^k\mathbf{v}||_2 \leq ||\mathbf{v}||_2$$

Moreover, we obtain the weak inequality on the eigenvalues of $\mathbf{W}$:

$$||\mathbf{Wv}||_2 \leq ||\mathbf{QWv}||_2 \leq ||\mathbf{v}||_2$$

The weak inequality shows the eigenvalues of $\mathbf{W}$ lie (weakly) within the unit circle. Since the eigenvalues of $\mathbf{W}^k$ are k-powers of the eigenvalues of $\mathbf{W}$, we see by the strong inequality that no eigenvalue can lie precisely on the unit circle. ∎

Lemma B.1.2. *Under Assumption 3.1.1, the beliefs of the agents, $\boldsymbol{\pi}_t$, converge almost surely to some $\boldsymbol{\pi}_\infty$.*

Proof. Fix $\delta > 0$. Recall that $\boldsymbol{\pi}_t = \boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_t) + \mathbf{W}\boldsymbol{\pi}_{t-1}$ for the DeGroot agents and we can treat stubborn agents as DeGroots with $\theta_i = 1$ and $\gamma_i = 0$. Notice by induction one can show that $\mathbf{0} \leq \boldsymbol{\pi}_t \leq \mathbf{1}$ (it is a belief): because $\boldsymbol{\pi}_0 = q\mathbf{1}$ and every belief update is a convex combination of $\text{BU}(\mathbf{h}_t)$, which lies between $\mathbf{0}$ and $\mathbf{1}$, and neighboring beliefs in the period $t-1$, which by the inductive hypothesis lie between $\mathbf{0}$ and $\mathbf{1}$, $\boldsymbol{\pi}_t$ must lie between $\mathbf{0}$ and $\mathbf{1}$. Moreover, by

Lemma 3.1.1, $\mathrm{BU}(\mathbf{h}_t)$ converges to a constant vector almost surely. Now let us write:

$$\begin{aligned}
||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 &= ||\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_t) - (\mathbf{I} - \mathbf{W})\boldsymbol{\pi}_{t-1}||_2 \\
&= ||\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_t) - (\mathbf{I} - \mathbf{W})(\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_{t-1}) + \mathbf{W}\boldsymbol{\pi}_{t-2})||_2 \\
&= ||\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_t) - (\mathbf{I} - \mathbf{W})(\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_{t-1}) + \mathbf{W}(\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_{t-2}) + \mathbf{W}\boldsymbol{\pi}_{t-3}))||_2 \\
&= ||\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_t) - (\mathbf{I} - \mathbf{W})(\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_{t-1})) - (\mathbf{I} - \mathbf{W})\mathbf{W}(\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_{t-2})) - (\mathbf{I} - \mathbf{W})\mathbf{W}^2\boldsymbol{\pi}_{t-3}||_2
\end{aligned}$$

Repeating this, we see that for any $t \geq M - 2$:

$$||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 = \left|\left|\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_t) - \boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W}) \sum_{k=0}^{M} (\mathbf{W}^k \mathrm{BU}(\mathbf{h}_{t-k-1})) - (\mathbf{I} - \mathbf{W})\mathbf{W}^{M+1}\boldsymbol{\pi}_{t-M-2}\right|\right|_2$$

Because $\mathbf{W}$ has a spectral radius which is strictly less than 1 by Lemma B.1.1, we know that $\lim_{k\to\infty} \mathbf{W}^k = \mathbf{0}$. Moreover, since $\boldsymbol{\pi}_t$ is bounded between $\mathbf{0}$ and $\mathbf{1}$, we know there exists some M^* such that $||(\mathbf{I} - \mathbf{W})\mathbf{W}^{M^*+1}\boldsymbol{\pi}_{t-M^*-2}||_2 \leq \frac{\delta}{3}$ and $||\boldsymbol{\theta} \odot \mathbf{W}^{M^*+1}\mathbf{1}||_2 \leq \frac{\delta}{3}$. Thus, for this value of M^* and any $t \geq M^* - 2$:

$$||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 \leq \left|\left|\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_t) - \boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W}) \sum_{k=0}^{M^*} \mathbf{W}^k \mathrm{BU}(\mathbf{h}_{t-k-1})\right|\right|_2 + \frac{\delta}{3}$$

Because $\mathrm{BU}(\mathbf{h}_t)$ converges to a constant almost surely by Lemma 3.1.1, we know there exists T^* almost surely such that for all $t > T^*$, $||\boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W})(\mathrm{BU}(\mathbf{h}_t) - \mathrm{BU}(\mathbf{h}_{t-k-1}))||_2 < \frac{\delta}{3(M^*+1)}$ for all $0 \leq k \leq M^*$ by the Cauchy criterion of convergence. Thus, for all $t > \max\{M^* - 2, T^*\}$:

$$\begin{aligned}
||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 &\leq \left|\left|\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_t) - \boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W}) \sum_{k=0}^{M} (\mathbf{W}^k \mathrm{BU}(\mathbf{h}_t))\right|\right|_2 + \frac{2\delta}{3} \\
&< \left|\left|\boldsymbol{\theta} \odot \mathrm{BU}(\mathbf{h}_t) - \boldsymbol{\theta} \odot (\mathbf{I} - \mathbf{W}^{M^*+1})\mathrm{BU}(\mathbf{h}_t)\right|\right|_2 + \frac{2\delta}{3} \\
&\leq \left|\left|\boldsymbol{\theta} \odot \mathbf{W}^{M^*+1}\mathbf{1}\right|\right|_2 + \frac{2\delta}{3}
\end{aligned}$$

(Note that because the spectral radius of $\mathbf{W}$ is less than 1 by Lemma B.1.1, $\sum_{k=0}^{M} \mathbf{W}^k = (\mathbf{I} - \mathbf{W})^{-1}(\mathbf{I} - \mathbf{W}^{M+1})$.) Recall we chose M^* such that the first term in the last expression does not exceed $\delta/3$. Thus, $||\boldsymbol{\pi}_t - \boldsymbol{\pi}_{t-1}||_2 < \delta$ for all $t > \max\{M^* - 2, T^*\}$, which completes the proof. ∎

Proof of Theorem 3.1.1. By Lemma 3.1.1 and Lemma B.1.2 we know that both $\mathrm{BU}(\mathbf{h}_t)$ and $\boldsymbol{\pi}_t$

converge almost surely to $\text{BU}(\mathbf{h}_\infty)$ and $\boldsymbol{\pi}_\infty$, respecitlvely. Thus, $\boldsymbol{\pi}_\infty$ must solve the fixed-point problem:

$$\boldsymbol{\pi}_\infty = \boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_\infty) + \mathbf{W}\boldsymbol{\pi}_\infty$$

If not, then the difference between the left-hand side and right-hand side is always some positive amount η, and so every iteration of belief updating changes the belief by at least η, contradicting convergence. By Lemma B.1.1, we know that all eigenvalues of $\mathbf{W}$ lie within the unit circle, so $\mathbf{I} - \mathbf{W}$ is invertible, and thus we can solve this fixed-point problem explicitly:

$$\boldsymbol{\pi}_\infty = (\mathbf{I} - \mathbf{W})^{-1}(\boldsymbol{\theta} \odot \text{BU}(\mathbf{h}_\infty))$$

which proves the claim of Proposition 3.1.1. ∎

Proof of Proposition 3.1.1. Whenever $p_i > 1/2$, by Lemma 3.1.1, the personal Bayesian update component (BU) of the DeGroot update converges almost surely to belief 1 on the true state, so $\text{BU}_i(h_{i,\infty}(\mathbf{0}))(R|S) \overset{a.s.}{\to} 0$. On the other hand, when $p_i = 1/2$ we have $\theta_i = 0$ by Assumption 3.1.1, so $\text{BU}_i(\mathbf{h}_\infty(\mathbf{0}))(y'|y) \odot \boldsymbol{\theta} \overset{a.s.}{\to} 0$. This implies that $\text{BU}(\mathbf{h}_\infty(\mathbf{0}))(y'|y) \odot \boldsymbol{\theta} \overset{a.s.}{\to} \mathbf{0}$ trivially. By Proposition 3.1.1, we see that for R:

$$\begin{aligned}\boldsymbol{\pi}_t(R) &\overset{a.s.}{\to} (\mathbf{I} - \mathbf{W})^{-1}(\text{BU}(\mathbf{h}_\infty(\mathbf{0}))(R) \odot \boldsymbol{\theta}) \\ &= (\mathbf{I} - \mathbf{W})^{-1}\mathbf{0} \\ &= \mathbf{0}\end{aligned}$$

Thus, $\boldsymbol{\pi}_t(S) \overset{a.s.}{\to} \mathbf{1}$, and agents learn the true state almost surely. ∎

Proof of Theorem 3.1.2. Suppose that agent i has belief $\pi_{i,T}(R)$, so agent i's best response is the action $a_i = R$ if $\pi_i(R) > (1-b)/2$, $a_i = S$ if $\pi_i(S) < (1-b)/2$, or any strategy in the simplex $\Delta(\{S, R\})$ if $\pi_i(R) = (1-b)/2$. Therefore, the action of the agents in the terminal stage is pinned-down as a function of terminal beliefs.

By Lemma B.1.2, as $T \to \infty$, the beliefs of all agents converge almost surely to some $\boldsymbol{\pi}_\infty$, given a network action $\mathbf{x}$. We can construct a set $\mathcal{B}$ which consists of all the values of b where some agent i has a limit belief $\lim_{t\to\infty} \pi_{i,t} \overset{a.s.}{\to} (1-b)/2$, for some network action $\mathbf{x}$. Note there is only one such b value per agent, given by $1 - 2\pi_{i,\infty}$. Thus, provided there are finitely many agents and finitely many principal influence actions, the set $\mathcal{B}$ is finite, so has measure zero, implying that $(-1,1)\backslash\mathcal{B}$ has full measure. Moreover, every agent either picks the correct terminal action or the incorrect terminal action, almost surely, for all $b \in (-1,1)\backslash\mathcal{B}$.

Consider fixing some $\mathbf{x}$ and any $b \in (-1,1)\backslash\mathcal{B}$. Given fixed ζ, for every $\kappa > 0$, there exists T^* such that for all $T > T^*$, the probability that all beliefs at time T are within ζ of their limits is at least $1-\kappa$:

$$\mathbb{P}[||\boldsymbol{\pi}_T - \boldsymbol{\pi}_\infty||_\infty < \zeta] \geq 1-\kappa$$

by Lemma B.1.2. Since the set of $\mathcal{B}$ contains no b's with an agent holding $\pi_{i,\infty} = (1-b)/2$, we can pick T^* large enough and ζ small enough whereby each agent i plays a known action a_i with probability at least $1-\kappa$ at time T. Choosing action $\mathbf{x}$ gives the principal a known net payoff of $k_1 - \varepsilon||\mathbf{x}||_1$ with probability $1-\kappa$ (which we deem the "likely payoff") and some other payoff with probability κ, where k_1 is the number of manipulated agents under (pure) strategy $\mathbf{x}$.

Now suppose two network strategies $\mathbf{x}_1, \mathbf{x}_2$ have a different number of manipulated agents, k_1 and k_2, respectively. If $\mathbf{x}_1$ and $\mathbf{x}_2$ give the same likely payoff, this implies that $k_1 - \varepsilon||\mathbf{x}_1||_1 = k_2 - \varepsilon||\mathbf{x}_2||_1$, which implies that:

$$\varepsilon = \frac{k_1 - k_2}{||\mathbf{x}_1||_1 - ||\mathbf{x}_2||_1}$$

because $||\mathbf{x}_1||_1 \neq ||\mathbf{x}_2||_1$. Noting that both the numerator and denominator are integers, we see that by taking the generic set of irrational ε, we guarantee that whenever $\mathbf{x}_1$ and $\mathbf{x}_2$ have a different number of manipulated agents, the principal has a strictly higher likely payoff under one. Since we took κ to be arbitrary, we can choose κ small (by increasing T) such that the principal prefers action $\mathbf{x}_1$ to $\mathbf{x}_2$ if he prefers the likely payoff of $\mathbf{x}_1$ to the likely payoff of $\mathbf{x}_2$ (as the expected payoff contribution of any "unlikely" payoff is bounded above by $n \cdot \kappa$, and $\kappa \to 0$). Thus, for the set of irrational ε and $b \in (-1,1)\backslash\mathcal{B}$, the principal plays the strategy over network actions which induces the "likely" outcome of that network action with probability at least $1-\kappa$. In such a strategy, the number of manipulated agents then must be the same, and the all of the principal's optimal strategies are manipulation-invariant. ∎

Proof of Proposition 3.1.2. We prove by induction that $\mathbf{W}_{ij}^{\ell}$ represents the sum of weighted walks of length ℓ between i and j, not passing through a stubborn agent. The base case of $\ell = 0$ is clear because every agent has a walk of length 0 to themselves of weight 1, and none others.

Note that:

$$\begin{aligned}
\mathbf{W}_{ij}^{\ell+1} = [\mathbf{W}\mathbf{W}^{\ell}]_{ij} &= \sum_{k=1}^{n} w_{ik}\mathbf{W}_{kj}^{\ell} \\
&= \sum_{k=1}^{n} w_{ik} \cdot [\text{weight of walks of length } \ell \text{ between } k \text{ and } j] \\
&= [\text{weight of walks of length } \ell+1 \text{ between } i \text{ and } j]
\end{aligned}$$

Therefore, the total weight of walks between i and j (avoiding stubborn agents) is given by $\sum_{W\in\mathcal{W}_{ij}} w_W = \sum_{\ell=0}^{\infty} \mathbf{W}^{\ell} = (\mathbf{I}-\mathbf{W})^{-1}$ since the spectral radius of $\mathbf{W}$ is strictly less than 1, by Lemma B.1.1. Finally, note that by Proposition 3.1.1:

$$\begin{aligned}
\pi_{i,\infty}(\mathbf{x}^*) &= (\mathbf{I}-\mathbf{W})_i^{-1}(\boldsymbol{\gamma}(\mathbf{x}^*)\odot\boldsymbol{\theta}) \\
&= \sum_{j=1}^{n} (\mathbf{I}-\mathbf{W})_{ij}^{-1}\gamma_j(\mathbf{x}^*)\theta_j \\
&= \sum_{j=1}^{n} \gamma_j(\mathbf{x}^*)\theta_j \left(\sum_{W\in\mathcal{W}_{ij}} w_W\right) \\
&= \mathcal{D}_i(\boldsymbol{\gamma})
\end{aligned}$$

As this holds for every i, we have $\mathcal{D}(\boldsymbol{\gamma}) = (\mathbf{I}-\mathbf{W})^{-1}(\boldsymbol{\gamma}\odot\boldsymbol{\theta})$. ∎

Section 4

Proof of Theorem 3.1.3. For part (a), we note that by Proposition 3.1.1, limiting DeGroot beliefs of the incorrect state R for $\boldsymbol{\theta} = \theta'\mathbf{1}$ are given by:

$$\boldsymbol{\pi}_{\infty}(R) = (\mathbf{I}-\mathbf{W}_{\theta'}^{-1})(\boldsymbol{\gamma}\odot\boldsymbol{\theta}')$$

We first prove that the asymptotic bound for DeGroot beliefs is continuous in θ' around $\theta' = 0$. Clearly the network preservation of $\mathbf{W}_{\theta'}$ is continuous in θ', so it is sufficient to prove that as $\theta' \to 0$, $\mathbf{I}-\mathbf{W}_{\theta'}$ is non-singular. To see this, note the eigenvalues of $\mathbf{W}_{\theta'}$ are uniformly bounded away from the unit circle as $\theta' \to 0$ (and thus $\mathbf{I}-\mathbf{W}_{\theta'}$ is non-singular as $\theta' \to 0$), so one can apply the same reasoning as Lemma B.1.1, noting that the existence of at least one stubborn agent guarantees $\mathbf{W}$ is still substochastic. Thus, provided $(\mathbf{I}-\mathbf{W}_{\theta'})^{-1}$ is a continuous operation at $\theta' = 0$, we can substitute $\theta' = 0$ and apply DeGroot centrality with influence vector $\boldsymbol{\gamma} \leq \mathbf{1}$, showing that all DeGroot centralities tend to 0, so beliefs of the correct state tend toward 1. This yields the claim in (a).

Because $\lim_{\theta' \to 1} \mathbf{W}_{\theta'} = \mathbf{0}$, it is obvious that beliefs are continuous at $\theta' = 1$. Moreover, when $\theta' = 1$, any DeGroot agent i is manipulated if and only if $\gamma_i = 1$, which is profitable if and only if $\varepsilon < 1$. Call the strategy of targeting all DeGroots as $\mathbf{1}_D$, which has a net utility of $(1-\varepsilon)(n-m)$. If $b < 1/2$, then (c) holds vacuously; to show (b), we just note by continuity that there exists some θ^{**} such that the network with $\theta' \in (\theta^{**}, 1)$ is either impervious (if $\varepsilon < 1$) or susceptible (if $\varepsilon > 1$) independent of θ'. Setting $\theta^* = \theta^{**}$ and $\overline{\theta} = (1+\theta^{**})/2$ gives us (b).

Now consider $b > 1/2$ and let $\theta^* = 1/2$. Suppose the principal chooses $\mathbf{1}_D$ with the only difference being that he does not target the DeGroot agent not adjacent to any stubborn agents; call this strategy $\mathbf{x}_{\text{spec}}$. By just considering first-order walks, we see that the DeGroot centrality of this agent is at least $(1-\theta^*)\theta^* = 1/4$, so this agent is still manipulated under $\mathbf{x}_{\text{spec}}$. Similarly since all other DeGroot agents *are* targeted and have $\theta = 1/2$, these agents are also manipulated. Therefore the net utility of strategy $\mathbf{x}_{\text{spec}}$ is $(1-\varepsilon)(n-m)+\varepsilon$, which beats $\mathbf{1}_D$. Let $\overline{\theta}$ be the infimum of all $\theta > 1/2$ where agent i is manipulated if and only if $\gamma_i = 1$ for all i (call this proprty **Independence**); we know such an infimum exists because independence holds at $\theta' = 1$. We claim that for all $\theta' \in (\overline{\theta}, 1)$, independence holds. To see this, it is sufficient to show that if independence holds with some θ'_1, then independence holds for any $\theta'_2 > \theta'_1$. By way of contradiction, consider some the strategy $\mathbf{x}_2$ which violates independence with θ_2 by targeting all agents except agent i^* who is manipulated. This implies that for some DeGroot i^*, the sum of weighted walks to other DeGroots j with $\gamma_j = 1$ exceeds $(1-b)/2$ with θ_2, given that all other agents receive $\gamma_j = 1$ but agent i has $\gamma_i = 0$. However, the sum of weighted walks with θ_1 is *necessarily* larger, because $\alpha_{ij,1} > \alpha_{ij,2}$ for all i, j. Thus, $\mathbf{x}_2$ violates independence under θ'_1, a contradiction.

By construction, there exists some $\varepsilon^* > 1 - 1/n$ such that $\theta' \in (\theta^*, \overline{\theta})$ is susceptible (because $\mathbf{x}_{\text{spec}}$ dominates $\mathbf{0}$) but where $\mathbf{x}_D$ is dominated by $\mathbf{0}$. Also by our previous observation, for $\theta' \in (\overline{\theta}, 1)$, an agent is manipulated if and only if $\gamma_i = 1$, so the network is impervious if and only if $\varepsilon > 1$, which holds for ε^*. Therefore, these $\theta^*, \overline{\theta}$ satisfy (b) and (c). ∎

Proof of Proposition 3.1.3. We will appeal to the first part of Corollary B.1.1. Let $j_2^* \in D_2$ be the agent in D_2 adjacent to an agent $j_1^* \in D_1$. Now consider an arbirary agent $j \in D_1$. Since D_1 is strongly connected, there exists a walk between j and j_1^*, which implies there is also a walk from j to j_2^*; let us denote this walk by $W_{jj_2^*} = j \to v_1 \to \cdots \to v_k \to j_1^* \to j_2^*$. Suppose $\theta_1 \in [0, \overline{\theta})$ for some $\overline{\theta} < 1$. Let us write the weight of this walk explicitly as:

$$w_{jj_2^*} = \theta_2 \prod_{(v_i \to v_{i+1}) \in W_{jj_2^*}} (1-\theta_1)\alpha_{v_i\, v_{i+1}} > C_{jj_2^*} > 0$$

where the constant $C_{jj_2^*}$ does not depend on θ_1, as $\theta_1 < \overline{\theta}$. If we take $\bar{b} = 1 - 2\min_{j \in D_1} C_{jj_2^*} < 1$, then we see that for all $b > \bar{b}$, all $j \in D_1$ have DeGroot centrality $\mathcal{D}_j(\mathbf{1}_D) \geq w_{jj_2^*} \geq C_{jj_2^*} \geq (1-b)/2$. Thus, all agents in D_1 are manipulated when ε is sufficiently small, regardless of their θ_1, and in particular as $\theta_1 \to 0$. On the other hand, all agents in D_2 have $\theta_2 \geq \min_{j \in D_1} C_{jj_2^*}$, so by the same argument agents in D_2 are manipulated.

The second result is just a rephrasing of Theorem 3.1.3(a). ∎

Proof of Proposition 3.1.4. Let there be M manipulated agents under optimal strategy x with influence cost ε, so the principal has a payoff of $M - \varepsilon||\mathbf{x}||_1$. After we increase ε to ε', suppose the principal manipulates more agents; it necessarily must be the case that $||\mathbf{x}'|| > ||\mathbf{x}||$, otherwise x is strictly preferred to x′ for any influence cost, so cannot be optimal with ε'. But then, of course:

$$\begin{aligned}
M' - \varepsilon||\mathbf{x}'||_1 &= M' - \varepsilon'||\mathbf{x}'||_1 + (\varepsilon' - \varepsilon)||\mathbf{x}'||_1 \\
&\geq M - \varepsilon'||\mathbf{x}||_1 + (\varepsilon' - \varepsilon)||\mathbf{x}'||_1 \\
&= M - \varepsilon||\mathbf{x}'||_1 + \varepsilon'(||\mathbf{x}'||_1 - ||\mathbf{x}||_1) \\
&\geq M - \varepsilon||\mathbf{x}'||_1 + \varepsilon(||\mathbf{x}'||_1 - ||\mathbf{x}||_1) \\
&\geq M - \varepsilon||\mathbf{x}||_1
\end{aligned}$$

which contradicts the optimality of x when the influence cost is ε.

Note that the DeGroot centrality of the agents under the same strategy x does not depend on b, but the cutoff necessary to take the incorrect action is $(1-b)/2$, so is decreasing in b. Thus, the number of manipulated agents (for a fixed network strategy), k, is non-decreasing in b. Therefore, if there exists some strategy x where $M - \varepsilon||\mathbf{x}||_1 > 0$, then when b increases to b', we know $M' \geq M$, so the same strategy x yields $M' - \varepsilon||\mathbf{x}||_1 > 0$. By Corollary B.1.1, the network with $b' > b$ is susceptible. ∎

Proof of Example 3.1.1. First, consider the belief cutoff $\pi_{\text{cutoff}}(R) = 0.35$ and where the principal targets agents 1 and 3. Then beliefs of the incorrect state are given by:

$$\boldsymbol{\pi}(R) = \begin{pmatrix} 1 & -1/2 & 0 & 0 \\ -1/3 & 1 & -1/3 & 0 \\ 0 & -1/3 & 1 & -1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1/2 \\ 0 \\ 1/3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.692 \\ 0.385 \\ 0.462 \\ 0 \end{pmatrix},$$

and all three DeGroot agents are manipulated, yielding a payoff of $3 - 2\varepsilon > 0$. Targeting all three agents will also lead to these three agents being manipulated, but increases the cost with no additional benefit. Clearly, if the principal targets no one, then all beliefs of R will be 0, which yields no profit. Thus, the only potential for a better strategy would be if the principal can manipulate two or more agents by sending signals to only one:

1. *Send to agent 1 only*: Only agent 1 is manipulated.

$$\boldsymbol{\pi}(R) = \begin{pmatrix} 1 & -1/2 & 0 & 0 \\ -1/3 & 1 & -1/3 & 0 \\ 0 & -1/3 & 1 & -1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1/2 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.615 \\ 0.231 \\ 0.077 \\ 0 \end{pmatrix}$$

2. *Send to agent 2 only*: Only agent 2 is manipulated.

$$\boldsymbol{\pi}(R) = \begin{pmatrix} 1 & -1/2 & 0 & 0 \\ -1/3 & 1 & -1/3 & 0 \\ 0 & -1/3 & 1 & -1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1/3 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.231 \\ 0.462 \\ 0.154 \\ 0 \end{pmatrix}$$

3. *Send to agent 3 only*: Only agent 3 is manipulated.

$$\boldsymbol{\pi}(R) = \begin{pmatrix} 1 & -1/2 & 0 & 0 \\ -1/3 & 1 & -1/3 & 0 \\ 0 & -1/3 & 1 & -1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ 1/3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.077 \\ 0.154 \\ 0.385 \\ 0 \end{pmatrix}$$

Thus, the optimal strategy with $\pi_{\text{cutoff}} = 0.35$ is to target agents 1 and 3 and all agents are manipulated.

Now let us consider the case of $\pi_{\text{cutoff}} = 0.2$ (manipulation is easier). Once again consider the three cases from before. In the cases when only agent 1 is targeted, both agents 1 and 2 are manipulated; when only agent 2 is targeted, both agents 1 and 2 again end up manipulated; when agent 3 is targeted, only agent 3 is manipulated. Thus there is a strategy that obtains a payoff of $2 - \varepsilon > 3 - 2\varepsilon > 3 - 3\varepsilon$. Therefore, no strategy that targets more agents (even if all three agents are manipulated!) beats the strategy of targeting just agent 1 or agent 2. And as before, targeting no one leads to manipulation and gives a payoff of 0. Thus, the optimal strategy with $\pi_{\text{cutoff}} = 0.2$ is to target just one agent and manipulate only two. ∎

Section 5

Proof of Theorem 3.1.4. This follows immediately from Proposition B.1.1, as noted in Appendix B.1.1, by taking $I_1 = \{1, \ldots, n\}$. ∎

Proof of Proposition 3.1.5. Suppose we sprinkle m stubborn agents such that $\lceil n/m \rceil$ is the farthest distance between any two "adjacent" stubborn agents along the ring. Then for all DeGroots i, letting $j^*(i)$ be the nearest stubborn agent (looking backward):

$$\begin{aligned} \mathcal{D}_i(\mathbf{1}_D) &= 1 - \prod_{\ell = j^*(i)+1}^{i} \left(1 - \frac{1}{n+1}\right) \\ &\leq 1 - \left(\frac{n}{n+1}\right)^{\lceil n/m \rceil} \\ &\leq 1 - e^{-2/m} \end{aligned}$$

For any b and $\boldsymbol{\gamma}$, we have that $\mathcal{D}_i(\boldsymbol{\gamma}) \leq \mathcal{D}_i(\mathbf{1}_D) \leq (1-b)/2$ if $m \geq \frac{2}{\log\left(\frac{2}{1+b}\right)}$, which as we see does not depend on n. Thus setting the constant $m^* = \frac{2}{\log\left(\frac{2}{1+b}\right)}$ obtains the claim. ∎

Proof of Proposition 3.1.6. Because θ is constant in n, when the principal plays $\boldsymbol{\gamma} = \mathbf{1}_D$, the DeGroot centrality of all agents depends only on their distance from the nearest stubborn agent $j^*(i)$, and not the population size n:

$$\begin{aligned}\mathcal{D}_i(\mathbf{1}_D) &= 1 - \prod_{\ell=j^*(i)}^{i} \frac{1}{2} \\ &= 1 - \frac{1}{2^{d(i,j^*(i))}}\end{aligned}$$

where $d(i, j^*(i))$ is the distance between agent i and (stubborn) agent $j^*(i)$. Thus, every DeGroot agent is manipulated if and only if she is (at least) a distance d^* away from her previous stubborn agent. Because a DeGroot agent is manipulated only if $\mathcal{D}_i(\mathbf{1}_D) > (1-b)/2$, we see that $d^* = 1 + \lceil \log_2\left(\frac{1}{1+b}\right)\rceil$.

Clearly, by setting $c = 1$, the network is impervious with $c \cdot n$ stubborn agents because all the agents are stubborn. On the other hand, when $c = 0$ and $\varepsilon < 1$, the principal makes positive utility by targeting every agent in the population and manipulating (almost) everyone, so by Corollary B.1.1, the network is susceptible. Now consider the infimum of all c such that the network with n agents remains impervious with some configuration of $\lfloor c \cdot n \rfloor$ stubborn agents. Call this value c^*, and by the previous two observations, we know that it exists and $c^* \in (0, 1)$.

We first show that $\lfloor c^* \cdot n \rfloor$ stubborn agents makes the network impervious. To do this, we establish that $c^* \cdot n$ is integral. If $c^* \cdot n$ is not integral, then the network with n agents is still impervious with $\lfloor c^* \cdot n \rfloor < c^* \cdot n$ agents, so is impervious with $c^{**} \cdot n$ agents where $c^{**} < c^*$, contradicting the definition of c^*. Thus, $c^* \cdot n$ is integral. Then, it is easy to see $c = c^* + \epsilon$ for small ϵ attains the same manipulation as with c^*, so the network is impervious with $c^* \cdot n$ stubborn agents.

By the definition of c^*, any fewer stubborn agents that $c^* \cdot n$ must make the network susceptible. With a non-sprinkled configuration, consider $\bar{d}_{\text{sprinkled}}$ and $\bar{d}_{\text{non}}$, the maximum distance from a stubborn agent for any DeGroot agent in the sprinkled and non-sprinkled configurations, respectively. By definition of "sprinkled," $\bar{d}_{\text{sprinkled}} < \bar{d}_{\text{non}}$. We claim that some DeGroot agent i on the chain between two stubborn agents which attains a distance of $\bar{d}_{\text{non}}$ must be manipulated. If not, the network is impervious when all agents are at a distance (less than or equal to) $\bar{d}_{\text{non}}$ from the last stubborn agent, as the principal employs identical strategies on identical length chains between two stubborn agents, by symmetry. Note that $\bar{d}_{\text{sprinkled}} = \lceil 1/c \rceil - 1$ and $\bar{d}_{\text{non}} \geq \bar{d}_{\text{sprinkled}} + 1$. Thus, the network is impervious with $c^{**} \cdot n$ stubborn agents, where c^{**} is the largest number such that $c^{**} \cdot n$ is integral and $\lceil 1/c^{**} \rceil = \lceil 1/c^{*} \rceil + 1$ (which is guaranteed to exist for large n). Clearly $c^{**} < c^{*}$, a contradiction of the definition of c^{*}. Thus, the non-sprinkled configuration is susceptible.

To see that $c^{*} \in \Theta(1)$, note that if $n_1 = kn_2$ for $k \in \mathbb{N}$ and $c \cdot n_1$ is integral, then the network with $c \cdot n_1$ stubborn agents is impervious to manipulation if and only if the network with $c \cdot n_2$ stubborn agents is impervious. This is immediate from the fact that any path of length z between two stubborn agents with n_1 agents in the population along the ring can be transformed into k paths of length z between two stubborn with n_2 agents along the ring. Again, because of symmetry, the principal must employ identical strategies on all k copies of the z-length path. ∎

Proof of Theorem 3.1.5. We can order the agents by their location on the ring, starting from some arbitrary agent 1. Fix the principal's influence vector $\boldsymbol{\gamma}$. We can write the DeGroot centrality of (DeGroot) agent i in network $\mathbf{G}_{\eta}$ as:

$$\mathcal{D}_i(\boldsymbol{\gamma}) = \left(\frac{\eta}{n+1} + \frac{1-\eta}{2}\right)\gamma_i + \frac{1-\eta}{2}\mathcal{D}_{i-1}(\boldsymbol{\gamma}) + \sum_{j=1}^{n} \frac{\eta}{n+1}\mathcal{D}_j(\boldsymbol{\gamma})$$

Summing over both sides, we obtain:

$$\begin{aligned}\sum_{i=1}^{n} \mathcal{D}_i(\boldsymbol{\gamma}) &= \left(\frac{\eta}{n+1} + \frac{1-\eta}{2}\right)||\boldsymbol{\gamma}||_1 + \frac{1-\eta}{2}\sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma}) + \frac{\eta(n-m)}{n+1}\sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma}) \\ &= \frac{(1-\eta)n + (1+\eta)}{2(n+1)}||\boldsymbol{\gamma}||_1 + \frac{\eta(n-1-2m) + (n+1)}{2(n+1)}\sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma})\end{aligned}$$

This gives us

$$\frac{(n+1) - \eta(n-2m-1)}{2(n+1)}\sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma}) = \frac{(1-\eta)n + (1+\eta)}{2(n+1)}||\boldsymbol{\gamma}||_1 \Longrightarrow \sum_{j=1}^{n}\mathcal{D}_j(\boldsymbol{\gamma}) = \frac{(1-\eta)n + (1+\eta)}{(n+1) - \eta(n-2m-1)}||\boldsymbol{\gamma}|$$

Let us call $\zeta(\boldsymbol{\gamma}) \equiv \frac{(1-\eta)n+(1+\eta)}{(n+1)-\eta(n-2m-1)}||\boldsymbol{\gamma}||_1$. If stubborn agents form a continuous chain or are there are only $o(n)$ many, then there exists a continuous chain in the ring of DeGroots that grows unboundedly in n (without any stubborn agents agents along the chain). Let agent i^* be the first DeGroot on such a chain. If the principal targets all agents along this chain, then:

$$\mathcal{D}_{i^*}(\boldsymbol{\gamma}) = \frac{\eta}{n+1} + \frac{1-\eta}{2} + \frac{\eta}{n+1}\zeta(\boldsymbol{\gamma})$$
$$\mathcal{D}_{i}(\boldsymbol{\gamma}) = \frac{\eta}{n+1} + \frac{1-\eta}{2} + \frac{1-\eta}{2}\mathcal{D}_{i-1}(\boldsymbol{\gamma}) + \frac{\eta}{n+1}\zeta(\boldsymbol{\gamma})$$

Solving the recursion, for an agent at location τ away from i^*, we see that:

$$\begin{aligned}\mathcal{D}_{\tau}(\boldsymbol{\gamma}) &= \left(\frac{(1-\eta)n+(1+\eta)}{2(n+1)} + \frac{\eta}{n+1}\zeta(\boldsymbol{\gamma})\right)\sum_{\tau'=0}^{\tau-1}\left(\frac{1-\eta}{2}\right)^{\tau'} \\ &= \left(\frac{(1-\eta)n+(1+\eta)}{2(n+1)} + \frac{\eta}{n+1}\zeta(\boldsymbol{\gamma})\right)\frac{1-((1-\eta)/2)^{\tau}}{1-(1-\eta)/2} \\ &\overset{\tau\to\infty}{\to} \frac{(1-\eta)n+(1+\eta)}{(n+1)(1+\eta)} + \frac{2\eta}{(n+1)(1+\eta)}\zeta(\boldsymbol{\gamma})\end{aligned}$$

It is easy to verify that $\mathcal{D}_{\tau}(\boldsymbol{\gamma})$ is decreasing in η. When $\eta = 0$, the principal can obtain a payoff that grows unboundedly in n by manipulating $\omega(1)$ agents along this chain of DeGroots, and no strategy that manipulates only $O(1)$ agents does better; therefore, $\omega(1)$ agents are manipulated. This reasoning continues to hold as long as $\mathcal{D}_{\tau}(\boldsymbol{\gamma}) \geq (1-b)/2$, and since $\boldsymbol{\gamma} = \mathbf{1}_D$ is profitable (given $\varepsilon < 1$), the condition $\mathcal{D}_{\tau}(\mathbf{1}_D) \geq (1-b)/2$ is both necessary and sufficient for imperviousness. Finally, by monotonicity and continuity of $\mathcal{D}_{\tau}(\boldsymbol{\gamma})$, we are guaranteed there exists η^* such that $\mathcal{D}_{\tau}(\boldsymbol{\gamma}) > (1-b)/2$ when $\eta < \eta^*$ and $\mathcal{D}_{\tau}(\boldsymbol{\gamma}) < (1-b)/2$ when $\eta > \eta^*$. ∎

Proof of Theorem 3.1.6. Because the stubborn agents are placed symmetrically and the network is symmetric itself, we know that $\mathcal{D}_i(\boldsymbol{\gamma}) = \mathcal{D}_j(\boldsymbol{\gamma})$ for all DeGroots $i, j \in D$. By definition, there exists ϕkm DeGroot-stubborn connections in the network. Once again, by symmetry, all (DeGroot) agents are adjacent to the same number of stubborn agents, m_*. We can compute m_* by computing the average connections to stubborn agents:

$$m_* = \frac{\phi km}{k(n-m)} = \phi\frac{m}{n-m}$$

By the recursive definition of DeGroot centrality, we see that:

$$\begin{aligned}
\mathcal{D}(\mathbf{1}_D) &= \frac{1}{1+k} + \frac{k}{1+k} \cdot \left(1 - \phi \frac{m}{n-m}\right) \mathcal{D}(\mathbf{1}_D) \\
&= \frac{1}{1+k} + \frac{k}{1+k} \frac{n-(1+\phi)m}{n-m} \mathcal{D}(\mathbf{1}_D) \\
\implies \quad \mathcal{D}(\mathbf{1}_D) &= \frac{n-m}{(\phi k - 1)m + n}
\end{aligned}$$

Simply rearranging with the observation that an agent is manipulated with $\mathcal{D}(\mathbf{1}_D) \leq (1-b)/2$, we see that if the principal plays $\mathbf{1}_D$, there is no manipulation if and only if $\phi kn/(n-m) \geq (1+b)/(1-b)$. Since $\mathcal{D}(\boldsymbol{\gamma}) \leq \mathcal{D}(\mathbf{1}_D)$ for all $\boldsymbol{\gamma}$, we see the network is impervious when this inequality holds. ∎

Section 6

Proof of Proposition 3.1.7. Consider the learning dynamics given by $\boldsymbol{\mu}_{t+1} = \tilde{\theta}_i \cdot \boldsymbol{\mu}_0 + \tilde{\mathbf{W}} \boldsymbol{\mu}_t$. By the same reasoning as in Theorem 3.1.1, we see that:

$$\boldsymbol{\mu}_t \overset{a.s.}{\to} (\mathbf{I} - \tilde{\mathbf{W}})^{-1} (\boldsymbol{\mu}_0 \odot \tilde{\boldsymbol{\theta}}) \equiv \boldsymbol{\mu}_\infty$$

Thus, as $T \to \infty$, it is sufficient to consider the learning dynamics given by:

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\mu}_\infty \cdot \boldsymbol{\theta} \odot \mathrm{BU}_i(h_{i,t+1}) + (\mathbf{1} - \boldsymbol{\theta} \odot \boldsymbol{\mu}_\infty) \oslash (\mathbf{1} - \boldsymbol{\theta}) \odot \mathbf{W} \boldsymbol{\pi}_t$$

where $\oslash$ is element-wise division. This is equivalent to the original learning dynamics, under a network preservation (see Definition 3.1.3) with $\boldsymbol{\theta}' = \boldsymbol{\mu}_\infty \cdot \boldsymbol{\theta}$. Plugging in the expression for $\boldsymbol{\mu}_\infty$, combined with the asymptotic beliefs given in Theorem 3.1.1, obtains the result. ∎

Proof of Proposition 3.1.8. Consider the case of concave cost with $X^* \geq \bar{X}$. Let M^*, M^{**} be the number of manipulated agents in the linear and concave cost cases, respectively, and X^*, X^{**} the number of targeted agents in the linear and concave cost cases, respectively. If $M^{**} < M^*$ (so $X^{**} < X^*$), then consider the payoff in the linear cost case from implementing the concave cost strategy:

$$\begin{aligned}M^{**}-\varepsilon X^{**} &= M^{*}-\varepsilon X^{*}+(M^{**}-M^{*})-\varepsilon(X^{**}-X^{*})\\ &\geq M^{*}-\varepsilon X^{*}+(M^{**}-M^{*})-(C(X^{**})-C(X^{*}))\\ &\geq M^{*}-\varepsilon X^{*}\end{aligned}$$

where the first inequality follows from the fact that $C(\cdot)$ is concave and $C(X^*) \geq C(\bar{X})$, and the second follows from the assumptions on M^{**}, M^*. But this contradicts the optimality of (M^*, X^*) in the linear cost case.

Now consider convex costs with $\bar{X}^* \geq \bar{X}$. Let M^*, M^{**} be the number of manipulated agents in the linear and convex cost cases, respectively, and X^*, X^{**} the number of targeted agents in the linear and convex cost cases, respectively. If $M^{**} > M^*$ (so $X^{**} > X^*$), then consider the payoff in the convex cost case from implementing the linear cost strategy:

$$\begin{aligned}M^{*}-\varepsilon X^{*} &= M^{**}-\varepsilon X^{**}+(M^{*}-M^{**})-(C(X^{*})-C(X^{**}))\\ &\geq M^{**}-\varepsilon X^{**}+(M^{*}-M^{**})-\varepsilon(X^{*}-X^{**})\\ &\geq M^{**}-\varepsilon X^{**}\end{aligned}$$

This contradicts the optimality of (M^{**}, X^{**}) in the convex cost case. Finally, note that in the convex cost case it is always true that when $X^* < \bar{X}$ that $C(X^*) < \varepsilon X^*$. Therefore, the strategy in the linear cost case necessarily obtains positive payoff for the principal, which means it improves on $\mathbf{x} = \mathbf{0}$, and so by Corollary B.1.1 the network is susceptible. ∎

Proof of Proposition 3.1.9. Let $\mathcal{D}_i^k(\boldsymbol{\gamma})$ and $\mathcal{D}_i(\boldsymbol{\gamma})$ denote the DeGroot centrality in k-cut subnetwork and the original network, respectively, under $\boldsymbol{\gamma}$. Suppose we have a k-cut subnetwork that is impervious to manipulation, so for any network strategy $\mathbf{x}$, we have that $\sum_{i\in D} \mathbf{1}_{\mathcal{D}^k(\boldsymbol{\gamma}(\mathbf{x}))>(1-b)/2} - \varepsilon x_i \leq 0$. Because $\varepsilon_u = 0$, it is sufficient to consider strategies $\mathbf{x}$ with $\gamma_u = 1$, as they dominate the strategies with $\gamma_u = 0$.

Consider the principal applying strategy $\mathbf{x}$ in the original network. First, we show the DeGroot centrality of every agent in the k-cut subnetwork ($\mathcal{D}_i^k$) is at least that in the original network ($\mathcal{D}_i$). In the k-cut subnetwork, since $\gamma_u = 1$, we have that $\mathcal{D}_u^k(\boldsymbol{\gamma}) = 1$ which is clearly an upper bound on all $\mathcal{D}_v(\boldsymbol{\gamma})$ for $v \in \mathcal{K}$ in the original network. Consider the recursive definition of DeGroot centrality, in both the k-cut subnetwork and the original network:

$$\mathcal{D}_i(\boldsymbol{\gamma}) = \theta_i\gamma_i + \sum_{j=1}^{n} \mathcal{D}_j(\boldsymbol{\gamma})$$
$$\mathcal{D}_i^k(\boldsymbol{\gamma}) = \theta_i\gamma_i + \sum_{j=1}^{n-k} \mathcal{D}_j^k(\boldsymbol{\gamma}) + \mathcal{D}_u^k(\boldsymbol{\gamma})$$

which admits a unique fixed point. Note the above is an increasing map in $\{\mathcal{D}_j(\boldsymbol{\gamma})\}_{j=1}^n$, and since $\alpha_{iu} = \sum_{j\in\mathcal{K}} \alpha_{ij}$ with $\mathcal{D}_u^k(\boldsymbol{\gamma}) \geq \mathcal{D}_u(\boldsymbol{\gamma})$, the fixed point $\{\mathcal{D}_j(\boldsymbol{\gamma})\}_{j=1}^n$ in relation to the fixed point $\{\mathcal{D}_j^k(\boldsymbol{\gamma})\}_{j=1}^{n-k} \cup \mathcal{D}_u^k(\boldsymbol{\gamma})$ must satisfy $\mathcal{D}_j(\boldsymbol{\gamma}) \leq \mathcal{D}_j^k(\boldsymbol{\gamma})$ for all $j \in \{1, \ldots, n-k\}$. Therefore, for all $\mathbf{x}$:

$$\sum_{i\in D} \mathbf{1}_{\mathcal{D}(\boldsymbol{\gamma}(\mathbf{x}))>(1-b)/2} - \varepsilon x_i \leq \sum_{i\in D} \mathbf{1}_{\mathcal{D}^k(\boldsymbol{\gamma}(\mathbf{x}))>(1-b)/2} - \varepsilon x_i \leq 0$$

which means $\mathbf{x} = \mathbf{0}$ is optimal, and there is no manipulation in the original network. ∎

Proof of Corollary 3.1.1. The local density result of Proposition B.1.1 guarantees that the k-cut subnetwork is impervious to manipulation (with the exception of vertex u) when the log-diameter condition is met. Then applying Proposition 3.1.9 shows that the original network is k-impervious. ∎

Proof of Proposition 3.1.10. By Lemma 3.1.1, $\text{BU}_i(S|h_{i,t})$ converges to 1 if $z_{i,t}^S - z_{i,t}^R \to \infty$ (or $z_{i,t}^R / z_{i,t}^S \to 0$) and it converges to 0 if $z_{i,t}^R - z_{i,t}^S \to \infty$ (or $z_{i,t}^S / z_{i,t}^R \to 0$). As the DeGroot agents update their beliefs mechanically, any strategy where the principal sends mixed messages (i.e., $\hat{y}_i \neq R$) is dominated by one where he sends $\hat{y}_i = R$. Note that if the principal sends messages at average intensity λ_i^*, then her signal distribution is given by:

$$\mathbb{P}[s_{i,t} = R|\theta = S] = \frac{\lambda_i^*}{\lambda + \lambda_i^*} + \frac{\lambda}{\lambda + \lambda_i^*}(1 - p_i)$$

which is greater than $1/2$ (so $z_{i,t}^S / z_{i,t}^R \to 0$) when $\lambda_i^* > \lambda(2p_i - 1)$ and less than $1/2$ (so $z_{i,t}^R / z_{i,t}^S \to 0$) when $\lambda_i^* < \lambda(2p_i - 1)$. Note that since $\tilde{\varepsilon}$ is continuous, the difference in average cost between $\lambda_i^* = \lambda(2p_i - 1) - \delta$ and $\lambda_i^* = \lambda(2p_i - 1) + \delta$ shrinks to 0 when $\delta \to 0$, so the principal maximizes her payoff by other choosing an average intensity of $\lambda_i^* = \lambda(2p_i - 1) + \delta$ for vanishing $\delta \to 0$, or $\lambda_i^* = 0$. Moreover, since $\tilde{\varepsilon}$ is convex, the optimal targeting policy that minimizes cost but obtains an average targeting intensity λ_i^* is the constant function $\lambda_i^*(t) = \lambda_i^*$. This obtains exactly an average cost of $\tilde{\varepsilon}(\lambda_i^*)$. ∎

B.1.3 Worked Examples

Demonstration of DeGroot Centrality

Example B.1.1 (Illustration of DeGroot Centrality)**.** Consider the triangle network in Figure B-1, with one stubborn agent and two DeGroot agents all talking to each other. Suppose the DeGroot agents listen to themselves and their friends equally so that $\theta_i = \alpha_{ij} = 1/3$ for $j \neq i$, as shown by the solid lines. Using Theorem 3.1.1, we can characterize the limiting beliefs of the DeGroots about the incorrect state $y' \neq y$:

$$\boldsymbol{\pi}(y') \overset{a.s.}{\to} \left(\mathbf{I} - \begin{pmatrix} 0 & 0 & 0 \\ 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 0 \end{pmatrix}\right)^{-1} \left(\begin{pmatrix} 0 \\ x_2 \\ x_3 \end{pmatrix} \odot \begin{pmatrix} 1 \\ 1/3 \\ 1/3 \end{pmatrix}\right)$$

$$= \begin{pmatrix} 0 \\ \frac{3}{8}x_2 + \frac{1}{8}x_3 \\ \frac{1}{8}x_2 + \frac{3}{8}x_2 \end{pmatrix}$$

To measure DeGroot centrality, let us first consider the stubborn-avoiding weighted walks from a DeGroot agent i back to itself. There is a unique such walk of length $2r$ for $r = 0, 1, 2, \ldots$ from i to i, with weight $(1/3)^{2r}$ (by simply pinging back and forth between the two DeGroots, as in the dashed lines). Therefore, the total weight of stubborn-avoiding walks from i to i is $\sum_{r=0}^{\infty}(1/3)^{2r} = \frac{9}{8}$. Similarly, there is a unique stubborn-avoiding walk of length $2r+1$ for $r = 0, 1, 2, \ldots$ from i to $j \neq i$, with weight $(1/3)^{2r+1}$. Therefore, the total weight of walks from i

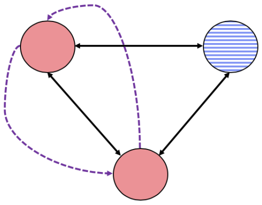


Figure B-1. Triangle Network (shaded agent = Stubborn; solid agents = DeGroot). Solid lines represent social network connections while dashed lines represent weighted walks that avoid stubborn agents.

to j is $\sum_{r=0}^{\infty}(1/3)^{2r+1} = \frac{3}{8}$. Using Definition 3.1.2, we see that for DeGroot i:

$$\begin{aligned}\mathcal{D}_i(\gamma) &= \frac{9}{8}\theta_i\gamma_i + \frac{3}{8}\theta_j\gamma_j \\ &= \frac{3}{8}x_i + \frac{1}{8}x_j\end{aligned}$$

which is equal to her belief of the incorrect state (as anticipated by Proposition 3.1.2). Thus in the above network, if the principal targets both DeGroots, their common belief in the incorrect state will be equal to $1/2$. Note that we get the same results if we instead use the recursive definition of DeGroot centrality. When $x_2 = x_3 = x$, we obtain by symmetry:

$$\begin{aligned}\mathcal{D}_i(\gamma) &= \frac{1}{3}\cdot x + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot \mathcal{D}_i(\gamma) \\ \implies \frac{2}{3}\mathcal{D}_i(\gamma) &= \frac{1}{3}x \\ \implies \mathcal{D}_i(\gamma) &= \frac{1}{2}x\end{aligned}$$

which coincides with the previous calculation when $x_2 = x_3 = x \in \{0, 1\}$. ∎

Applications of Theorem 3.1.4

Example B.1.2 (Complete Network)**.** Consider the complete network on n vertices. We suppose that, for simplicity, $\theta_i = \alpha_{ij} = 1/(n+1)$ for all DeGroot agents i and agents j (of any kind). This corresponds to each agent weighing each source of opinion (each neighbor, plus their own news) equally. The log-diameter of this network is exactly $\log(n+1)$ for any $n \geq 2$. Therefore, only a constant number of stubborn agents are needed by Theorem 3.1.4 (applying the result for $\delta = 1$), and in particular, one can show that $m \geq (1+b)/(1-b)$ are required for the complete network of size n. ∎

Example B.1.3 (Influential Star Network)**.** Consider Figure B-2 which shows one type of star network. We suppose that, for simplicity, $\theta_i = 1/(n+1)$ for all agents; that is, each agent weighs its own news as if it were in the complete network. Let agent 1 be the central agent of the star and agents $\{2, \dots, n\}$ be on the periphery. For agent $i \in \{2, \dots, n\}$, we have $\alpha_{i1} = n/(n+1)$ and $\alpha_{ij} = 0$ for all other j. For agent 1, we have $\alpha_{1j} = 1/(n+1)$ for all agents j. In other words, the central agent is *highly influential*, as all peripheral agents are influenced much more by this

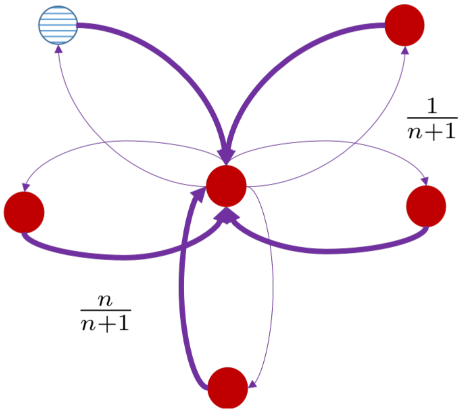


Figure B-2. Influential Star Network. A weighted directed arrow from node i to node j indicates that i puts that much weight on j's belief. Shaded node represents stubborn agents.

agent than their own news.

Once again, for any $n \geq 2$, the log-diameter of the network is at most $\log(n+3)$; between any two agents on the periphery, we have $\log((n+1)^2/n) = \log(n+2+1/n) \leq \log(n+3)$. In fact, if the number of stubborn agents satisfies $m \geq 2(1+b)/(1-b)$, the network is impervious. This is true even when all of the stubborn agents are on the periphery. So, in a seemingly very asymmetric network, still only a constant number are needed.

This does not imply, however, that fewer stubborn agents would not be sufficient to make the network impervious, if placed in better positions. For instance, a single stubborn agent in the center of the star *always* makes the network impervious when n is large enough. To see this, we can apply the local density result of Proposition B.1.1,[1] by considering subsets $I_\ell = \{1, \ell\}$ for $\ell = 2, \ldots, n$. The log-distance of each I_ℓ is given by $\log(1+1/n) = \log(|I_\ell|+1/n-1) \leq \log(|I_\ell|+1)$. Thus, when $b = 0$ and applying the bound in Example B.1.2, we see that if the stubborn agent is the central agent, the network is impervious (whereas we would require $m \geq 2$ on the periphery). ∎

Example B.1.4 (Echo chambers)**.** Consider Figure B-3, with a clique of size $n-1$ consisting entirely of DeGroot agents and a single stubborn agent. We assume the clique and the stubborn agents are joined by just a single (bidirectional) link connecting the stubborn agents with one DeGroot in the clique, and with no other connections going between the islands. We call this DeGroot agent the *pivotal agent*. This defines an undirected social network $\mathbf{G}^*$. For simplicity, let $\mathbf{G}$ have the weights given by $\theta_i = \alpha_{ij} = 1/(1+|N(i)|)$ whenever $i \to j$ in $\mathbf{G}^*$.

[1] We cannot apply Theorem 3.1.4 here because the stubborn agents disconnects all the DeGroots from each other, so the log-diameter is $+\infty$.

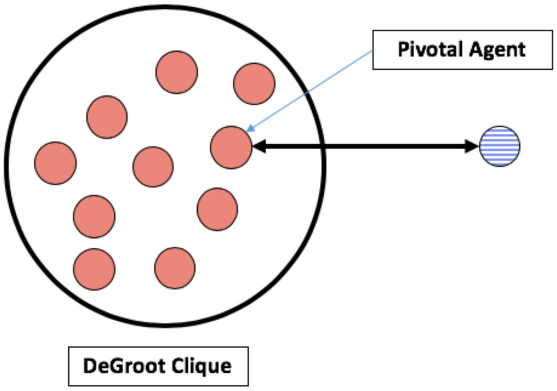


Figure B-3. A clique of DeGroot agents with a single connection to a stubborn agent.

We see that the pivotal agent can reach any other agent with a path of log-weight at most $-\log\left(\frac{1}{n}\right) = \log(n)$ and therefore satisfies the density condition with $\delta = 0$. Notice, however, that if the principal targets all agents in the DeGroot clique, then when n is large, the pivotal agent will still have an arbitrarily incorrect belief, i.e., $\pi_{i,T}(R) \to 1$ as $n \to \infty$. To see this, note that as $n \to \infty$, almost every walk (of any length) from the pivotal agent ends up at another DeGroot agent.[2] Because DeGroots only talk amongst themselves, there is an *echo chamber* whereby the misinformation sent by the principal circulates within the DeGroot island and the beliefs of the stubborn agents never propagate. Therefore, while the pivotal agent is close to the stubborn agent, the fact that most of her friends, and friends of friends are not, almost all of the influence exerted on the pivotal agent comes from others exposed to misinformation.

Compare this to the case where *every* DeGroot agent is pivotal, which now satisfies the log-diameter condition for $\delta = 1$. Even though DeGroot agents are friends almost exclusively with other DeGroot agents, who receive possible misinformation, Theorem 3.1.4 guarantees the network is impervious. This is precisely because an echo chamber effect no longer amplifies incorrect beliefs of the agents, simply because each DeGroot agent is friends with at least *one* stubborn agent, limiting the principal's influence. ∎

Finally, we expand on Example B.1.4 to show how Theorem 3.1.4 applies only to *log-diameter*, which may not coincide with the notion of diameter in undirected networks:

Example B.1.5 (Echo chambers, revisited)**.** Consider a variant of the unweighted social network

[2] To be precise, **G** has weights that represent a random walk for all DeGroot agents, where the agent chooses a link uniformly at random. The probability the walk ever reaches a stubborn agent is $\frac{1}{n}\left(1 + \left(\frac{n-1}{n}\right)\left(\frac{1}{n-1}\right)\sum_{k=0}^{\infty}\left(\frac{n-2}{n-1}\right)^k\right) = \frac{1}{n}\left(\frac{n^2+n-1}{n^2}\right)$, and as $n \to \infty$, tends toward 0.

$\mathbf{G}^*$ of the "echo chambers" network from Example B.1.4, but now where there are two cliques of size $n/2$, one clique which is all DeGroot and one clique which is all stubborn, with a single connection between them. Note that $\mathbf{G}^*$ has a diameter of 3, since it is possible to get from any agent in one clique to any other agent in the other clique with a walk that has no more than 3 steps. Because the diameter of the network stays constant with n, it is natural to classify this as a "small diameter" network.

Yet, straightforward computation reveals that the log-diameter of $\mathbf{G}$ is $\log\left(\frac{n^4}{2(n-1)^2}\right) \approx \log(n^2/2)$, which does not satisfy the conditions of Theorem 3.1.4 for any δ. In fact, as we saw before in Example B.1.4, no constant number of stubborn agents are guaranteed to make this network impervious. Therefore, having a small diameter in $\mathbf{G}^*$, even as n grows, does not necessarily imply the conditions of Theorem 3.1.4 will be satisfied for small log-diameter. ∎

Sparse Networks

Finally, the last sparse example is the *balanced star network*, where agents are aligned in a star network but employ equal-influence weighting. We show that despite the seemingly added symmetry, as compared to Example B.1.3, the network fails to satisfy the log-diameter condition, and so introduces unique vulnerabilities not present in the asymmetric star network of Example B.1.3.

Example B.1.6 (Balanced Star Network)**.** Consider the balanced star network of Figure B-4. Suppose that for agents on the periphery $\theta_i = \alpha_{i1} = 1/2$ whereas the core agent 1 updates as in Example B.1.3, $\theta_1 = \alpha_{1j} = 1/(n+1)$. The log-diameter condition is unsatisfied because the log-diameter grows as $\approx \log(2n)$.

When the central agent is stubborn, then either all of the agents are manipulated (if $b < 0$ and $\varepsilon < 1$) or none of them are (otherwise), i.e., the network is impervious. If stubborn agents are only on the periphery, then if $m \leq \beta n$ for all $\beta > 0$ as n grows large (i.e., the number of peripheral stubborn agents is sublinear), Stubborn agents have a vanishing fraction of influence in the network. The DeGroot centrality of the core agent converges to $\mathcal{D}_1(\boldsymbol{\gamma}) = ||\boldsymbol{\gamma}||_1/n$, whereas the DeGroot centrality of the peripheral agent i converges to $\mathcal{D}_i(\boldsymbol{\gamma}) = \frac{1}{2}\gamma_i + \frac{1}{2}||\boldsymbol{\gamma}||_1/n$. In other words, for peripheral agents, their belief is half of the average news experience and half of their own experience, whereas the core agent's belief is simply an average of all experiences.

Given a sublinear number of stubborn agents, the network is impervious if and only if

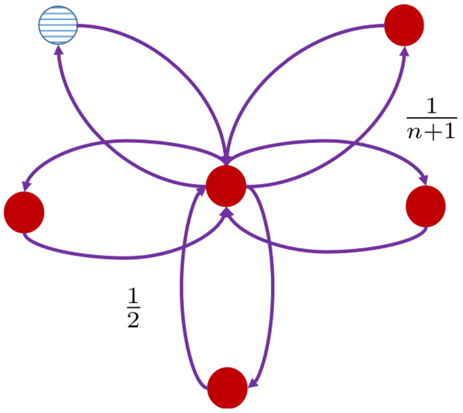


Figure B-4. Balanced Star Network

$\varepsilon < \max\{1/(1-b), 1\}$ for large n; otherwise, a *linear* number of stubborn agents on the periphery are required to prevent manipulation. If $b > 0$, then the principal targets $(1-b)$ fraction of the population; if $b < 0$, the principal targets all agents in the network, except the central agent. We note that the principal *targets the core agent last*, in contrast to the influential star network of Example B.1.3, where the principal should target this agent first. While the balanced star network is more symmetric in that no agent has disproportionate influence on the population, it also prevents the central agent from acting as a spokesperson for the knowledgable stubborn agents on the periphery. ∎

To conclude, we present an application of the results in Section 3.1.6 which allow us to characterize when a network is almost impervious but still has a few agents manipulated:

k-imperviousness

Example B.1.7. Consider the (unweighted) core-periphery network $\mathbf{G}^*$ shown in Figure B-5, with $n-k$ agents in the core and k agents on the periphery who listen to only one agent in the core (in Figure B-5, $k = 3$). Suppose the weights are given by the equal-influence weighting scheme. Fixing k, the log-diameter of the network is bounded below by $\log(3n)$ for sufficiently large n, which does not satisfy the conditions of Theorem 3.1.4 for any value of δ as n grows. On the other hand, there is an obvious k-cut which leaves the complete network as k-cut subnetwork (and after removing u), which has a log-diameter bounded above by $\log(n+1)$ (for sufficiently large n), and therefore is k-impervious with at least $m^*(k+1)$ stubborn agents located in the core via Example B.1.2 (the complete network is dense) and Corollary 3.1.1 (density condition for k-impervious). ∎

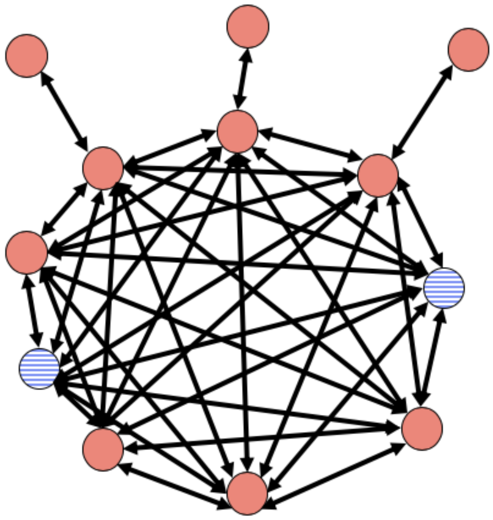


Figure B-5. Core-Periphery Network.

B.1.4 Numerical Experiments

The previous examples show how our results can be applied to the network topologies commonly studied in the literature. In this section, we examine these results in the context of real-world network data coming from Jackson et al. (2012). The network we consider represents an advice network in an Indian village, and consists of 144 nodes and 320 edges, where an edge between nodes i and j represents undirected communication between these two agents. In the following we look at different placement of stubborn agents in this network in order to further demonstrate the concepts introduced throughout the paper.

Similar to the setup we have so far, the principal tries to manipulate a subset of agents in the population by sending messages to some agents (not necessarily the same set of agents he is trying to manipulate) in the network. We compute the optimal strategy for the principal given the network topology (and we assume for simplicity that all weights θ are fixed at $\frac{1}{n}$). We start with Figure B-6 as an illustration that shows the network with only a single knowledgeable stubborn agent. Throughout the figures in this section, green nodes represent stubborn agents, and nodes represented with an asterisk indicate agents directly targeted by the principal (according to his *optimal* strategy). Conversely, DeGroot agents are colored either blue or red, to indicate whether under the principal's strategy the agent is manipulated (red) or not (blue). Thus, a network of all-blue and green agents means that this particular placement of the stubborn agents results in a network that is impervious to manipulation.

Throughout we fix $\varepsilon = 1/2$ (recall ε is the cost of sending messages to a single agent). For our first two examples, we consider the game in Table 3.1 and assume that $b = 0$, i.e. that agents' terminal actions reflect whichever state they believe is more likely. We focus on two particular agents, referred to in the data as Agent 70 and Agent 59. In Figure B-6, Agent 70 (with degree 7 and eigenvector centrality 0.0121) is a stubborn agent whose location results in no manipulation, because the principal has no profitable strategy with which he can manipulate even a single member of the population. Naturally, all agents have a DeGroot centrality of 0 when the principal chooses not to exert any influence.

On the other hand, Agent 59 is much more peripheral in the network, with a degree of 2 and eigenvector centrality of 0.0044. If Agent 59 is the stubborn agent, as is the case in Figure B-7, then the average DeGroot centrality (and terminal belief under the principal's optimal strategy) is $\bar{\pi} = 0.529 > \pi_{\text{cutoff}} \equiv 0.5$ and manipulation is inevitable and quite severe.

These two cases are summarized in Figure B-8. Each dot in this graph represents the DeGroot centrality of the corresponding agent in the network under one of the these two stubborn agent placements, and under a particular strategy for the principal:

1. *Optimal influence*: corresponds to the DeGroot centrality of the agents when the principal exerts the influence he would in his optimal strategy.

2. *Max influence*: corresponds to the DeGroot centrality of the agents when the principal targets every DeGroot agent, even though such influence may be "overkill" or ineffective.

Agents whose DeGroot centralities are above $\pi_{\text{cutoff}} = 0.5$ are manipulated. Yellow dots correspond to the DeGroot centality of the agents in Figure B-6 (with Agent 70), but when the principal employs max influence. Notice that all the yellow dots are below the threshold of π_{cutoff}, and hence no agent is manipulated despite the most intensive efforts of the principal. Thus, the stubborn agent communicates the truth effectively, and the principal cannot interfere. On the other hand, if the principal applies the same max-influence strategy to the network in Figure B-7 (with Agent 59 as the stubborn agent) then, as can be seen from the red dots, every single DeGroot agent is manipulated since all DeGroot centralities lie above the cutoff.

Most importantly in Figure B-8 however are the purple dots lying just above the dotted cutoff line, corresponding the principal's optimal strategy. These dots represent the DeGroot centralities of the agents in Figure B-7 when the principal applies the optimal targeting strategy

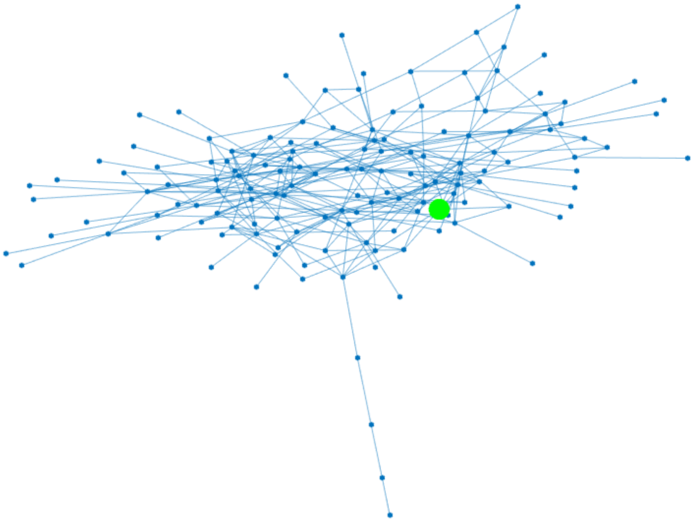


Figure B-6. Central Stubborn Agent, $b = 0$.

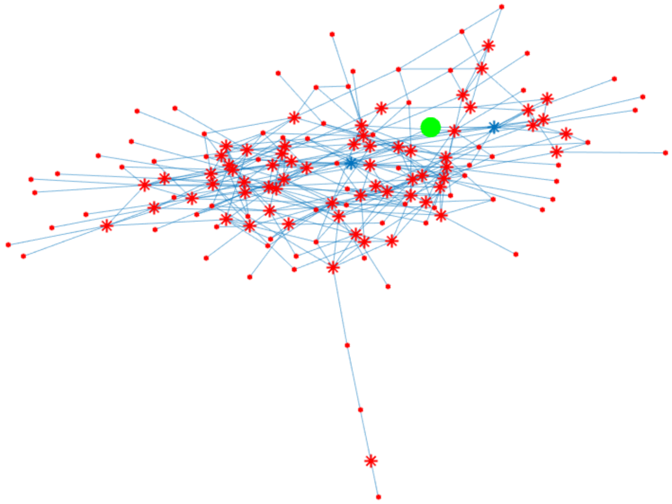


Figure B-7. Peripheral Stubborn Agent, $b = 0$.

depicted in the figure. Note that despite targeting 67 agents (46% of the population) instead of the entire population, the principal is able to obtain almost the maximum manipulation possible at a fraction of the cost (expends less than 50% of the cost), with only three agents (such as Agent 60 in the figure) escaping manipulation ($< 2\%$ of the population).

The rest of the figures examine the situation for different values of b. We have seen that when b is equal to zero, manipulation is very sensitive to the placement of the *single* stubborn agent. As b becomes lower and the cost of taking the risky action and mismatching the state increases, manipulation becomes exceedingly difficult. Similarly, as b increases, it becomes less costly for the agents to take the risky action, and hence it becomes easier to manipulate them. Figure B-9 shows that with $b = 0.5$, two stubborn agents (instead of one) are now required to prevent manipulation, provided they occupy network positions that again lead to low DeGroot centralities (across all γ) for the other agents. Similar to the ring network studied earlier, both the number and location of the stubborn agents matter. Figure B-10 shows that even with five stubborn agents agents, large-scale manipulation is possible because these agents occupy less central positions. In the case of the complete network, three stubborn agents are both necessary and sufficient for imperviousness when $b = 0.5$; in other words, the best-case placement in this network is better than in the complete network (requires only two stubborn agents) but the worst-case placement in this network is also worse than the complete network

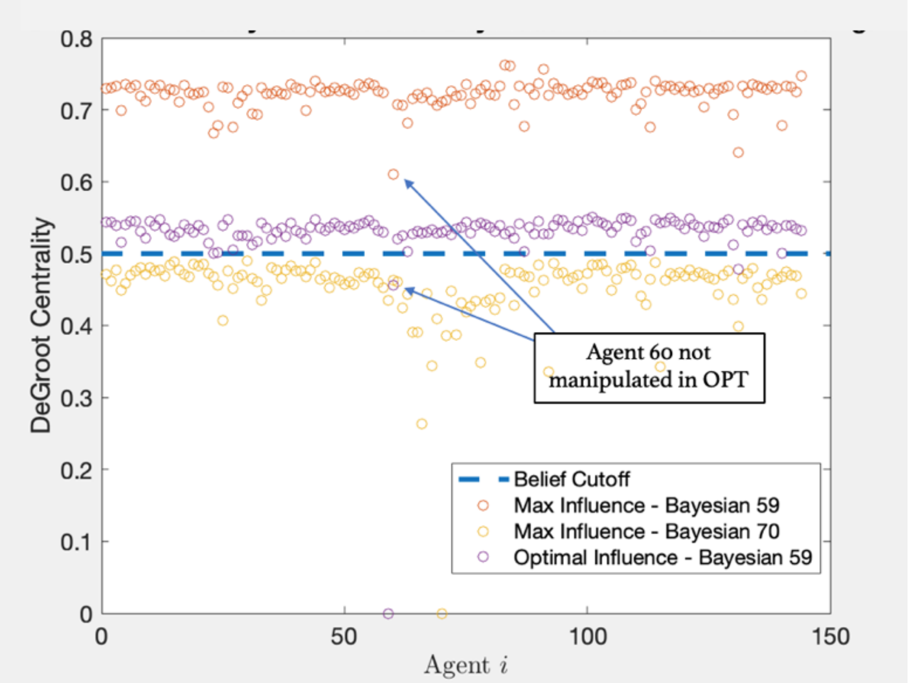


Figure B-8. DeGroot Centrality for Single Stubborn Agent, $b = 0$.

(requires at least six stubborn agents).

B.2 DeGroot Models II: Social Inequality and Misinformation

B.2.1 Technical Conditions and Model Details

Appendix B.2.1 provides more technical details about the deterministic model in Mostagir et al. (2022), while Appendix B.2.1 shows how to adapt that model to random networks with different communities and different levels of access to resources. Finally, Appendix B.2.1 demonstrates the main methods used in the proofs of this paper.

News Generation and Belief Evolution

The following model details are from Mostagir et al. (2022) and are presented here for contextualization of Section 3.3.1.

(a) **Organic News**: We assume agents receive organic information about the state y over time. News is generated according to a Poisson process with unknown parameter $\lambda_i > 0$ for each agent i; for simplicity, assume λ_i has atomless support over $(\underline{\lambda}, \infty)$ and $\underline{\lambda} > 0$. Let us denote

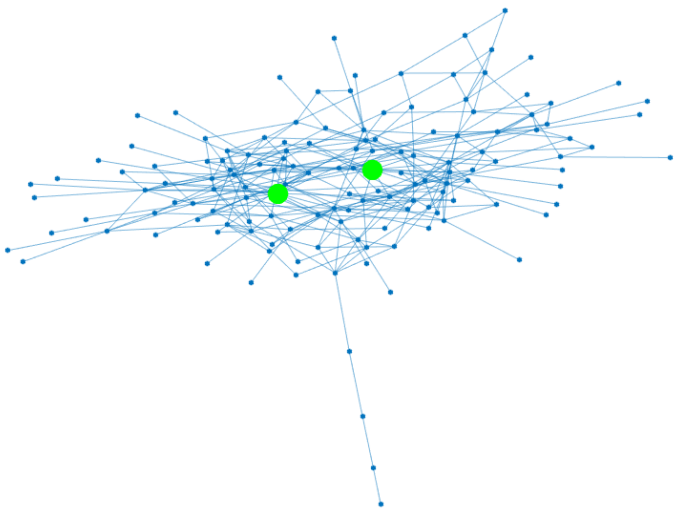


Figure B-9. Two Well-Placed Knowledgeable Stubborn Agents, $b = 0.5$.

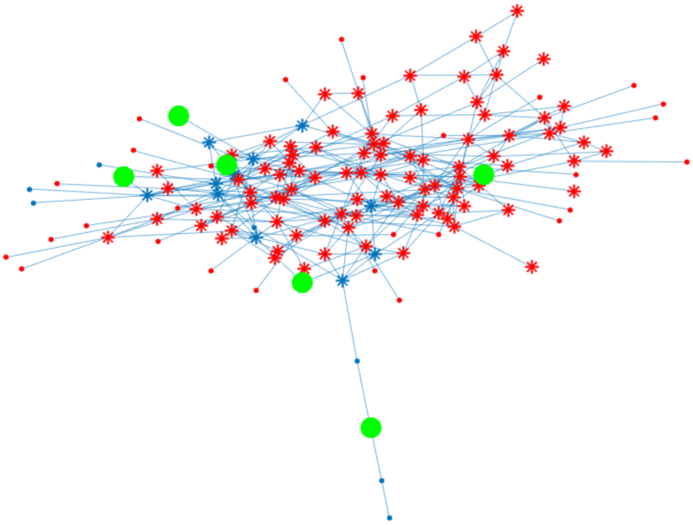


Figure B-10. Five Poorly-Placed Knowledgeable Stubborn Agents, $b = 0.5$.

by $(t_1^{(i)}, t_2^{(i)}, \ldots)$ the times at which news occurs for agent i. For all $\tau \in \{1, 2, \ldots\}$, the organic news for agent i generates a signal $s_{t_\tau^{(i)}} \in \{S, R\}$ according to the distribution:

$$\mathbb{P}\left(s_{t_\tau^{(i)}} = S \middle| y = S\right) = \mathbb{P}\left(s_{t_\tau^{(i)}} = R \middle| y = R\right) = p_i \in [1/2, 1)$$

i.e., the signal is correlated with the underlying truth.

(b) **News from Principal**: In addition to the organic news process, there is a principal who may also generate news of his own. At $t = 0$, the principal picks an influence state $\hat{y} \in \{S, R\}$. The principal then picks an influence strategy $x_i \in \{0, 1\}$ for each agent i in the network. If the principal chooses $x_i = 1$, for any agent i, then he (the principal) generates news according to an independent Poisson process with (possibly strategically chosen) intensity λ_i^* which is received by all agents where $x_i = 1$. We assume the principal commits to sending signals at this intensity, which may not exceed some threshold $\bar{\lambda}$.

(c) **News Observations**: Agents are unable to distinguish news sent by the principal or that organically generated. We denote by $\hat{t}_1^{(i)}, \hat{t}_2^{(i)}, \ldots$ the arrival of *all* news, either from organic sources or from the principal, for agent i. At each time $\hat{t}_\tau^{(i)}$, if the news is organic, the agent gets a signal according to the above distribution, whereas if the news is sent from the principal, she gets a signal of $\hat{y}$.

(d) **DeGroot Update**: DeGroots use a simple learning heuristic to update beliefs about the underlying state from other agents. We assume every DeGroot agent believes signals arrive according to a Poisson process and all signals are independent over time with $\mathbb{P}\left(s_{i,\hat{t}^{(i)}_\tau} = y\right) = p_i$ (i.e., takes the news at face value). DeGroot agents form their opinions about the state both through their own experience (i.e., the signals they receive) and the beliefs of their neighbors. Given history $h_{i,t} = (s_{i,\hat{t}^{(i)}_1}, s_{i,\hat{t}^{(i)}_2}, \ldots, s_{i,\hat{t}^{(i)}_{\tau_i}})$ up until time t with $\tau_i = \max\{\tau : \hat{t}^{(i)}_\tau \leq t\}$, each agent forms a personal belief about the state according to Bayes' rule. Let $z^S_{i,t}$ and $z^R_{i,t}$ denote the number of S and R signals, respectively, that agent i received by time t; then the DeGroot agent has a direct "personal experience":

$$\text{BU}(S|h_{i,t}) = \frac{p_i^{z^S_{i,t}}(1-p_i)^{z^R_{i,t}}q}{p_i^{z^S_{i,t}}(1-p_i)^{z^R_{i,t}}q + p_i^{z^R_{i,t}}(1-p_i)^{z^S_{i,t}}(1-q)}$$

As mentioned in Section 3.3.1, each DeGroot i then updates her belief for all $k\Delta < t \leq (k+1)\Delta$ according to:

$$\pi_{i,t} = \theta_i \text{BU}(h_{i,t}) + \sum_{i=1}^{n} \alpha_{ij}\pi_{j,k\Delta}$$

for some weights θ_i, α_{ij} with $\theta_i + \sum_{j=1}^n \alpha_{ij} = 1$, and Δ is a time period of short length. Note for simplicity in this paper we assume $\theta_i = \theta$ for all DeGroots i and $\alpha_{ij} = \frac{1-\theta}{|N(i)|}$ for all $j \in N(i)$.

(e) **DeGroot Centrality Vector**: To figure out the limit beliefs of the DeGroot agents when the principal targets everyone who is not a knowledgeable agent, denote by $\boldsymbol{\gamma}$ the vector in $\{0,1\}^n$ that designates which agents are targeted by the principal and let $\gamma_i = x_i = 1$ wherever agent i is DeGroot and $\gamma_i = 0$ everywhere else. DeGroot centrality, which is equivalent to the belief in the incorrect state in the limit is then given by $\mathcal{D}(\boldsymbol{\gamma}) = (\mathbf{I} - \mathbf{W})^{-1}\boldsymbol{\gamma} = \sum_{k=0}^{\infty} \mathbf{W}^k\boldsymbol{\gamma}$, where $\mathbf{I}$ is the identity and $\mathbf{W}$ is the adjacency matrix of weights given in Mostagir et al. (2022). DeGroot agent i is manipulated if her belief in the false state is above the cutoff, i.e. if $\mathcal{D}_i(\boldsymbol{\gamma}) > (1-b)/2$. More detailed methods on the computation of DeGroot centrality is given in B.2.1.

Model Adaptation to Stochastic-Block Networks

Using the proof technique in Mostagir and Siderius (2021), the following result establishes that as long as the population is large enough, it is sufficient to analyze the deterministic analogues of the random networks introduced in Section 3.3.1:

Theorem B.2.1. *For almost all b,*[3] *as $n \to \infty$, the probability that a random network drawn from the weak or strong inequality models has the same number of manipulated agents as the expected network converges to 1.*[4]

Theorem B.2.1 offers a technical simplification: instead of analyzing the random networks drawn according to the weak and strong homophily models, we can treat these networks as deterministic objects where the weights are chosen proportionally to the probability of link formation. Moreover, because all of our results are invariant to a doubling of the population as long as the proportion of knowledgeable and DeGroot agents remains constant on every island, every example can be extended to the case of $n \to \infty$ and Theorem B.2.1 can be applied.

In the original model of Mostagir et al. (2022), the authors derived results for arbitrary (but deterministic) network structures, where the principal must make (essentially) a binary decision for each agent whether to send her misinformed signals. In this paper, we instead adapt this model for random social networks which embed a notion of inequality in the form of unequal access to educational resources. At the core of the random network process are "islands" (or communities) where agents within an island have similar resources, but can have different resources from agents on different islands. Thus, the principal's optimal strategy instead is more subtle: he can target a fraction of the population on a given island (a rational number between 0 and 1). Because of symmetry, the principal does not care which agents on the island he actually targets, just the total percentage. Recall agents put θ weight on their own personal experience. This implies that, in the standard model, there will be two belief types within a given island, depending on whether the principal targets the agent directly or not.

To avoid this, and to add parsimony to the model, we assume agents will perform the personal-experience Bayesian update using the "average" news sent to the island. In particular, if the principal targets κ_ℓ fraction of the population on an island, the Bayesian update part of

[3] Recall that b is the parameter in **??**. "Almost all" is meant in the measure theoretic sense; the only exceptions lie on a set of measure 0 in $(-1, 1)$.

[4] The formal definitions of realized and expected networks can be found in Mostagir and Siderius (2021)

the belief update converges to $1 - \kappa_\ell$ as $T \to \infty$. This implies that all agents' beliefs on the same island will be the same, and allows us to study manipulation in the context of how certain communities are affected versus others.

Note this adaptation does not change much from the standard model. For instance, if θ is not too large (and agents rely significantly on social learning), then the differences in beliefs of two agents on the same island will be small in the standard setup. Thus, our results generalize easily to the case of the standard model as well, where agents are treated as individual binary decisions for the principal.

Finally, we note that all results implicitly assume that $n \to \infty$, because only under these conditions does Theorem B.2.1 equate the manipulation in random networks with that of their expected counterparts. Thus, while knowledgeable agent counts are technically discrete objects, because all of our results are closed under multiplication of the knowledgeable and DeGroot populations on each island by the same constant, we can think of "knowledgeable proportions" on each island and need not worry about whether such proportions divide the population size without remainder.

DeGroot Centrality: General Methods

Note that given a fixed strategy for the principal, the beliefs of (DeGroot) agents within a given island are the same due to symmetry as $n \to \infty$.[5] Here, we will introduce the general methodology for determining whether a population (or community) whose structure is randomly drawn from either the strong or weak homophily model is susceptible to manipulation. Let us define some notation:

(i) κ_ℓ is the fraction of island ℓ targeted by the principal (note that who he specifically targets on the island is immaterial);

(ii) $\mathcal{N}_\ell$ is the "neighborhood" of island ℓ; in the weak homophily model it is equal to $\{\ell' | \ell' \neq \ell\}$

[5] See Appendix B.2.1.

whereas in the strong homophily model it is equal to:

$$\begin{cases} \{2\}, \text{ if } \ell = 1 \\ \{\ell-1, \ell+1\}, \text{ if } \ell \in \{2, \ldots, k-1\} \\ \{\ell-1\}, \text{ if } \ell = k \end{cases}$$

One can compute DeGroot centralities by simply counting the weighted walks to misinformed agents, as in Figure B-11. However, DeGroot centrality computations are easiest when considering the linear recursive formulation of weighted walks, as in Figure B-12.

This calculation is done in two parts. The first part involves computing weighted walks to knowledgeable agents from agent i living on some island ℓ, which we denote as w_ℓ^K. The second part involves computing weighted walks to DeGroots who *do not directly* consume misinformation from the principal, which we denote as w_ℓ^D. The belief of the agent on island ℓ is then given by $w_\ell = w_\ell^K + w_\ell^D$.

We can calculate this explicitly as:

$$\frac{w_\ell^K}{1-\theta} = \frac{p_s m_\ell + \sum_{\ell' \in \mathcal{N}_\ell} p_d m_{\ell'}}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} + (1-\theta) \frac{p_s(s_\ell n - m_\ell)}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} w_\ell^K + (1-\theta) \frac{\sum_{\ell' \in \mathcal{N}_\ell} p_d (s_{\ell'} n - m_{\ell'})}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} w_{\ell'}^K$$

Similarly, we have:

$$\begin{aligned} \frac{w_\ell^D}{1-\theta} = \theta \frac{1-\kappa_\ell}{1-\theta} + \theta \frac{p_s(1-\kappa_\ell)(s_\ell n - m_\ell) + \sum_{\ell' \in \mathcal{N}_\ell} p_d (1-\kappa_{\ell'})(s_{\ell'} n - m_{\ell'})}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} \\ + (1-\theta) \frac{p_s(1-\kappa_\ell)(s_\ell n - m_\ell)}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} w_\ell^D + (1-\theta) \frac{\sum_{\ell' \in \mathcal{N}_\ell} p_d (1-\kappa_{\ell'})(s_{\ell'} n - m_{\ell'})}{p_s s_\ell n + \sum_{\ell' \in \mathcal{N}_\ell} p_d s_{\ell'} n} w_{\ell'}^D \end{aligned}$$

In particular, belief w_ℓ^K and w_ℓ^D both admit linear matrix equations with a closed-form solutions:

$$\begin{aligned} \frac{\mathbf{I}\mathbf{w}^K}{1-\theta} = \boldsymbol{a}^K + \boldsymbol{B}^K \mathbf{w} \implies \mathbf{w}^K = \left(\frac{\mathbf{I}}{1-\theta} - \boldsymbol{B}^K \right)^{-1} \boldsymbol{a}^K \\ \frac{\mathbf{I}\mathbf{w}^D}{1-\theta} = \boldsymbol{a}^D + \boldsymbol{B}^D \mathbf{w} \implies \mathbf{w}^D = \left(\frac{\mathbf{I}}{1-\theta} - \boldsymbol{B}^D \right)^{-1} \boldsymbol{a}^D \end{aligned}$$

where the total belief of the correct state is $w_\ell = w_\ell^K + w_\ell^D$, which is the complement of agent i on island ℓ's DeGroot centrality (i.e., $\mathcal{D}_\ell = 1 - w_\ell$).

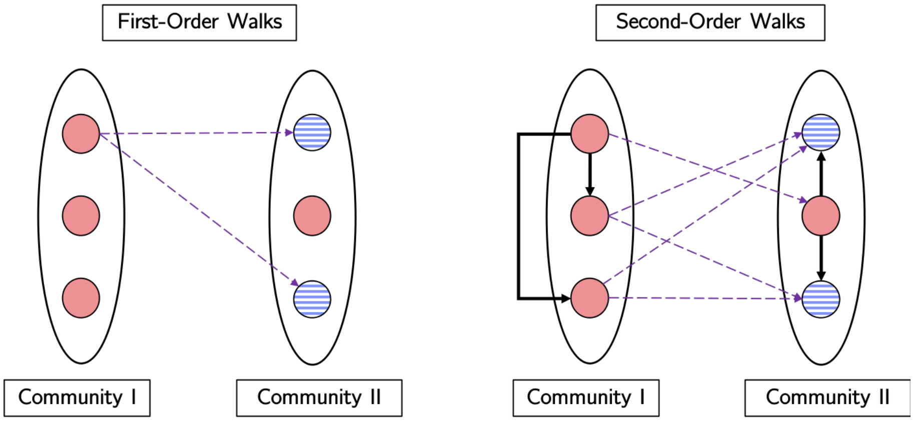


Figure B-11. An illustration of computing weighted walks to knowledgeable agents. Solid circles are DeGroot agents and shaded circles are knowledgeable agents. Solid lines represent higher weights "within-community links" than dashed lines. Consider the top-left agent, and for each walk, multiply the weights of the links along the walk. The figure on the left shows a first-order walk, i.e. a walk of length 1, which consists of the link directly connecting that agent to a knowledgeable agent. The second-order walk displayed on the right consists of walks of length 2, so that there is a link to another DeGroot agent who is linked to a knowledgeable agent, and the weight of that walk is the product of the two link weights and so on. Total weighted walks is the sum over all orders (i.e., walk lengths) of walks $1, 2, \ldots$.

B.2.2 Proofs

Preliminaries The following notation is used throughout the proofs. The vector $\boldsymbol{\gamma} \in \{0,1\}^n$ denotes which agents are targeted by the principal, and the DeGroot Centrality vector resulting from this targeting is denoted by $\mathcal{D}(\boldsymbol{\gamma})$. DeGroot agent on island ℓ is manipulated if $\mathcal{D}_\ell(\boldsymbol{\gamma}) > (1-b)/2$. Equivalently, we write π_ℓ or w_ℓ as the belief of an agent on island ℓ (the latter explicitly referring to walk counting, but is equivalent to π_ℓ).

Section 2

Proof of Theorem B.2.1. The proof uses results from Mostagir and Siderius (2021). The main result of that paper shows that one can focus on expected instead of random networks if three assumptions are satisfied. Assumption 1 from the Mostagir and Siderius (2021) holds because our networks are drawn from an inhomogeneous Erdos-Renyi model. Assumption 2 from that paper is also satisfied because θ is constant and the homophily models are connected almost surely. Similarly, the *expected degrees* and *normal society* conditions are satisfied because for the former, the expected degrees grow linearly in n for both the weak and

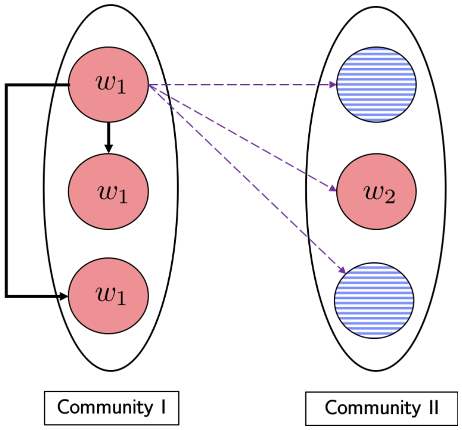


Figure B-12. An analytic approach to computing walks for the top-left agent (which equals her belief of the true state). Each agent's sum of walks equals a weighted-average of her neighbors' sums of walks.

strong inequality models, and for the latter, θ is the same for all agents. Therefore, we can apply Theorem 1 from the note for DeGroot centrality on an arbitrary targeting vector γ, i.e., $\lim_{n\to\infty}\mathbb{P}\left[||\tilde{\mathcal{D}}^{(n)}(\boldsymbol{\gamma}) - \bar{\mathcal{D}}^{(n)}(\boldsymbol{\gamma})||_\infty > \epsilon\right] = 0$. Thus, as $n \to \infty$:

$$\lim_{n\to\infty}\left(\mathbb{P}\left[\tilde{\mathcal{D}}_i^{(n)} < (1-b)/2 < \bar{\mathcal{D}}_i^{(n)} \text{ for some } i\right] + \mathbb{P}\left[\bar{\mathcal{D}}_i^{(n)} < (1-b)/2 < \tilde{\mathcal{D}}_i^{(n)} \text{ for some } i\right]\right) = 0$$

except for countably many b. Thus, for generic b, the number of manipulated agents is the same under both the expected and realized networks in the weak and strong inequality models, as $n \to \infty$. ∎

Section 4

Proof of Theorem 3.2.1. Notice the network is susceptible to manipulation (with high probability) if there exists an agent j with $\mathcal{D}_j(\mathbf{1}_D) > (1-b)/2$; similarly, the network is impervious to manipulation (with high probability) if for all agents j, $\mathcal{D}_j(\mathbf{1}_D) < (1-b)/2$. We show that increasing homophily leads to an increase in the inequality of DeGroot centralities (i.e., $\mathcal{D}(\mathbf{1}_D)$).

We have the system of equations:

$$\frac{1}{1-\theta}\mathbf{w} = \frac{k}{n(p_s+(k-1)p_d)}\begin{pmatrix} p_s m_1 + p_d \sum_{\ell\neq 1} m_\ell \\ p_s m_2 + p_d \sum_{\ell\neq 2} m_\ell \\ \dots \\ p_s m_k + p_d \sum_{\ell\neq k} m_\ell \end{pmatrix}$$

$$+ \frac{k(1-\theta)}{n(p_s+(k-1)p_d)}\begin{pmatrix} p_s(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ p_d(n/k-m_1) & p_s(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ \dots & \dots & \dots & \dots \\ p_d(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_s(n/k-m_k) \end{pmatrix}\mathbf{w}$$

which is equivalent to:

$$\frac{k}{(1-\theta)n(p_s+(k-1)p_d)}\mathbf{w} = \begin{pmatrix} p_s m_1 + p_d \sum_{\ell\neq 1} m_\ell \\ p_s m_2 + p_d \sum_{\ell\neq 2} m_\ell \\ \dots \\ p_s m_k + p_d \sum_{\ell\neq k} m_\ell \end{pmatrix}$$

$$+ (1-\theta)\begin{pmatrix} p_s(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ p_d(n/k-m_1) & p_s(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ \dots & \dots & \dots & \dots \\ p_d(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_s(n/k-m_k) \end{pmatrix}\mathbf{w}$$

Without loss of generality suppose that island k has the least number of knowledgeable agents of any island. Consider the map T given by:

$$T:\mathbf{w} \mapsto \begin{pmatrix} p_s m_1 + p_d \sum_{\ell\neq 1} m_\ell \\ p_s m_2 + p_d \sum_{\ell\neq 2} m_\ell \\ \dots \\ p_s m_k + p_d \sum_{\ell\neq k} m_\ell \end{pmatrix} + (1-\theta)\begin{pmatrix} p_s(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ p_d(n/k-m_1) & p_s(n/k-m_2) & \cdots & p_d(n/k-m_k) \\ \dots & \dots & \dots & \dots \\ p_d(n/k-m_1) & p_d(n/k-m_2) & \cdots & p_s(n/k-m_k) \end{pmatrix}\mathbf{w}$$

We claim that T has the property that $(w_\ell \geq w_k) \implies T(w_\ell) \geq T(w_k)$. Suppose that $w_\ell \geq w_k$, then:

$$\begin{aligned} m_\ell + (1-\theta)w_\ell(n/k - m_\ell) &\geq m_k + (1-\theta)w_\ell(n/k - m_k) \\ &\geq m_k + (1-\theta)w_k(n/k - m_k) \end{aligned}$$

which moreover implies that

$$\begin{aligned} & p_s(m_\ell + (1-\theta)w_\ell(n/k - m_\ell)) + p_d(m_k + (1-\theta)w_k(n/k - m_k)) + \sum_{\ell' \neq \ell, k} p_d(m_{\ell'} + (1-\theta)w_{\ell'}(n/k - m_{\ell'})) \\ \geq & p_d(m_\ell + (1-\theta)w_\ell(n/k - m_\ell)) + p_s(m_k + (1-\theta)w_k(n/k - m_k)) + \sum_{\ell' \neq \ell, k} p_d(m_{\ell'} + (1-\theta)w_{\ell'}(n/k - m_{\ell'})) \end{aligned}$$

because $p_s > p_d$. Because p_s, p_d, n are fixed, this suggests the map $\frac{k}{(1-\theta)n(p_s+(k-1)p_d)} \cdot T$ also has this property, so any fixed point of T must have $w_\ell \geq w_k$ by Brouwer's fixed point theorem for all islands ℓ. Since the system is linear and non-singular, there is a unique fixed-point with $w_\ell \geq w_k$. This implies the DeGroot centrality of the island with the least knowledgeable agents is always maximal and determines whether the network is impervious.

For the remainder of this part of the proof, we define a new operator T which maps $\mathbf{w}$, parametrized by p_s, p_d, and $\mathbf{m}$, respectively. We show the following: (i) $T|p_s$ is decreasing in p_s for w_k, (ii) $T|p_d$ is increasing in p_d for w_k, and (iii) $T|\mathbf{m}$ subject to $\sum m_\ell = m$ is increasing with every "Robin Hood" operation that puts more knowledgeable agents on island k for w_k.[6] This result suffices in order to show that there exists a fixed-point of T which obeys the desired properties of Theorem 3.2.1 (by Brouwer[7]), and by linearity, this fixed-point is unique.

[6] Note that other "Robin Hood" operations do not affect w_k, so does not affect the imperviousness of the network.

[7] In particular, let (w_1, w_2) be the old fixed-point and (w_1', w_2') the new fixed point. We illustrate for the case of increasing p_s: all other cases are similar. By increasing p_s, we know that T maps all w_1 larger and all w_2 smaller. Therefore, the convex compact set $[w_1, 1] \times [0, w_2]$ maps into itself, which implies the new fixed-point (w_1', w_2') lies in this set.

1. Decreasing in p_s: Let us define T as:

$$T|p_s : \mathbf{w} \mapsto \frac{1}{p_s + p_d}\Big[\begin{pmatrix} p_s m_1 + p_d \sum_{\ell \neq 1} m_\ell \\ p_s m_2 + p_d \sum_{\ell \neq 2} m_\ell \\ \dots \\ p_s m_k + p_d \sum_{\ell \neq k} m_\ell \end{pmatrix}$$
$$+ (1-\theta) \begin{pmatrix} p_s(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ p_d(n/k - m_1) & p_s(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ \dots & \dots & \dots & \dots \\ p_d(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_s(n/k - m_k) \end{pmatrix} \mathbf{w}\Big]$$

Computing directly:

$$\frac{\partial T(w_k|p_s)}{\partial p_s} = p_d \frac{(m_k + (1-\theta)(n/k - m_k)w_k) - \sum_{\ell \neq k} (m_\ell + (1-\theta)(n/k - m_\ell)w_\ell)}{(p_s + p_d)^2} < 0$$

where the inequalities follow from the analysis above.

2. Increasing p_d: Let us define T in the same way as in (1), except parametrized by p_d. Then in exactly the same way:

$$\frac{\partial T(w_k|p_d)}{\partial p_d} = -\frac{p_s}{p_d}\frac{\partial T(w_k|p_s)}{\partial p_s} > 0$$

which is the desired result.

3. Majorization: Assume that we remove a knowledgeable agent from island ℓ^* and add it to island k, with the assumption that $m_k + 1 \leq m_{\ell^*} - 1$. There are two cases: (i) island k still has the fewest knowledgeable agents (and thus the greatest DeGroot centrality), or (ii) some other island had the exact same number of knowledgeable agents as island k. In the latter case, the majorization does not affect whether the network is impervious or

susceptible. In the former case, define T as:

$$T|\mathbf{m} : \mathbf{w} \mapsto \Bigg[\begin{pmatrix} p_s m_1 + p_d \sum_{\ell \neq 1} m_\ell \\ p_s m_2 + p_d \sum_{\ell \neq 2} m_\ell \\ \cdots \\ p_s m_k + p_d \sum_{\ell \neq k} m_\ell \end{pmatrix}$$

$$+ (1-\theta) \begin{pmatrix} p_s(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ p_d(n/k - m_1) & p_s(n/k - m_2) & \cdots & p_d(n/k - m_k) \\ \cdots & \cdots & \cdots & \cdots \\ p_d(n/k - m_1) & p_d(n/k - m_2) & \cdots & p_s(n/k - m_k) \end{pmatrix} \mathbf{w} \Bigg]$$

Computing the directional derivative along the gradient $\mathbf{u} = \mathbf{e}_k - \mathbf{e}_{\ell^*}$ (where $\mathbf{e}_i$ is the vector of all 0's except for a 1 in ith spot):

$$\frac{\partial T(w_k)|\mathbf{m}}{\partial m_k} - \frac{\partial T(w_k)|\mathbf{m}}{\partial m_{\ell^*}} = p_d \left(1 - (1-\theta) w_k\right) - p_s \left(1 - (1-\theta) w_{\ell^*}\right)$$

Note that because $w_k \leq w_{\ell^*}$, $(1 - (1-\theta)w_k) \geq (1 - (1-\theta)w_{\ell^*})$. Because $p_s > p_d$, the above expression is positive.

Lastly, we need to argue that in the case of islands of equal size, increased DeGroot centrality inequality (i.e., the island with larger centrality increases its centrality while the other's centrality decreases) cannot make the network go from susceptible to impervious. To check if the network is impervious, all that needs to be checked is $\max_i \mathcal{D}_i(\mathbf{1}_D) > (1-b)/2$. When inequality of the DeGroot centrality increases, then $\max_i \mathcal{D}_i(\mathbf{1}_D)$ increases, and so the above inequality is more likely to be satisfied when inequality is increased. Therefore, the network can go from impervious to susceptible, but not the other direction. ∎

Proof of Theorem 3.2.2. Part (i) is a direct implication of Theorem 3.2.1: it is impossible for an increase in inequality to make the network switch from susceptible to impervious. Thus, if some inequality configuration makes the network impervious, it must necessarily be the case that the network with the least inequality is impervious.

For parts (ii) and (iii), consider the following construction of the inequality structures. If for all choices of $(\mathbf{m}, p_d)$ the DeGroot centralities of all of the islands are monotone in p_s (either monotonically increasing or decreasing) then choose p_s sufficiently close to 1 such that manipulation is the same under this inequality structure and $p_s = 1$ (i.e., extreme homophily). Such a $p_s < 1$ is guaranteed because centralities are continuous in the inequality parameters. Otherwise, there exists $(\mathbf{m}, p_d)$ such that at least one island has non-monotone centrality in p_s. First, note that all centralities are concave in p_s; this can be seen from considering the map in the proof of Theorem 3.2.1 for p_s and noting that:

$$\frac{\partial^2 T(w_\ell | p_s)}{\partial p_s^2} = \left(m_\ell + (1-\theta)(n/k - m_\ell) w_\ell - \sum_{\ell \neq k} (m_\ell + (1-\theta)(n/k - m_\ell) w_\ell) \right) \cdot \frac{p_s^2 - p_d^2}{(p_s + p_d)^4} > 0$$

because both terms in the above expression are positive. (Recall that w_ℓ is equal to 1 minus centrality, so convexity of w_ℓ corresponds to concavity of centrality.) Second, note that the centrality curve of an island with more knowledgeable agents always lies above an island with fewer (this was shown in Theorem 3.2.1). Suppose some island that exhibits the non-monotonicity of centrality in p_s; if there are multiple, pick the island whose centrality apex occurs at the largest value for p_s (call this is the "special" island); call this value p_s^*. If we choose b so that $(1-b)/2$ lies just below the apex of the centrality curve, then this island will be manipulated for p_s^*, but not at $p_s = 1$ because the centrality curve for the special island is concave (and thus is decreasing after p_s^*). All islands with more knowledgeable agents than the special island are protected when there is the most inequality, and all islands with fewer knowledgeable agents than the special island are manipulated under p_s^*. Similarly, every other island has either: (i) monotonically decreasing centrality, (ii) monotonically increasing centrality, or (iii) non-monotone centrality. In the case of (i) and (iii), by assumption, the island cannot be manipulated at $p_s = 1$ but is at p_s^*. Moreover for islands of type (ii), the centrality must naturally lie above the centrality curve of the special island, so is manipulated at p_s^*. Thus, this intermediate inequality structure has strictly more manipulation than the most inequality.

It just remains to show that for every $k \geq 3$, there exists a choice of m and distribution $\mathbf{m}$ that has at least one non-monotone centrality curve. For this, we provide an explicit example for $k = 3$. The parameters are given by $\theta = 1/20$, $p_d = .2$, $m_1 = 44$, $m_2 = 29$, $m_3 = 0$, $n = 1000$, for three islands. The plot is given in Figure B-13.

To generalize to more communities, simply add a community with all DeGroot agents. This will make the centrality of the special island increase for intermediate inequality relative to most inequality because the centrality of the all DeGroot island will exceed that of the special

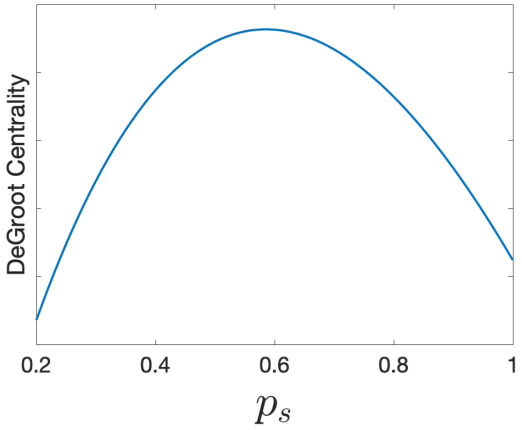


Figure B-13. DeGroot centrality for island 2 as a function of p_s.

island. Thus, there will still be manipulation with intermediate inequality, but not with the most inequality. ∎

Proof of Theorem 3.2.4. As in the proof of Theorem 3.2.1, let us consider each of the inequality cases separately and define corresponding maps T. We show that *all* beliefs decrease following an increase in p_s, decrease in p_d, or a reverse "Robin Hood" operation.

1. $\underline{p_s}$: Let us define T as:

$$T|p_s : \mathbf{w} \mapsto \left[\begin{pmatrix} \frac{p_s m_1 + p_d \sum_{\ell \neq 1} m_\ell}{p_s s_1 + (1-s_1) p_d} \\ \frac{p_s m_2 + p_d \sum_{\ell \neq 2} m_\ell}{p_s s_2 + (1-s_2) p_d} \\ \cdots \\ \frac{p_s m_k + p_d \sum_{\ell \neq k} m_\ell}{p_s s_k + (1-s_k) p_d} \end{pmatrix} + (1-\theta) \begin{pmatrix} \frac{p_s(s_1 n - m_1)}{p_s s_1 + (1-s_1) p_d} & \frac{p_d(s_2 n - m_2)}{p_s s_1 + (1-s_1) p_d} & \cdots & \frac{p_d(s_k n - m_k)}{p_s s_1 + (1-s_1) p_d} \\ \frac{p_d(s_1 n - m_1)}{p_s s_2 + (1-s_2) p_d} & \frac{p_s(s_2 n - m_2)}{p_s s_2 + (1-s_2) p_d} & \cdots & \frac{p_d(s_k n - m_k)}{p_s s_2 + (1-s_2) p_d} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{p_d(s_1 n - m_1)}{p_s s_k + (1-s_k) p_d} & \frac{p_d(s_2 n - m_2)}{p_s s_k + (1-s_k) p_d} & \cdots & \frac{p_s(s_k n - m_k)}{p_s s_k + (1-s_k) p_d} \end{pmatrix} \mathbf{w} \right]$$

Computing directly for island 1:

$$\begin{aligned} \frac{\partial T(w_1|p_s)}{\partial p_s} &= p_d \frac{(1-s_1)\left(m_1 + (1-\theta)(ns_1 - m_1)w_1\right) - s_1 \sum_{\ell \neq 1} \left(m_\ell + (1-\theta)(ns_\ell - m_\ell)w_\ell\right)}{(p_s s_1 + p_d(1-s_1))^2} \\ &= \frac{1}{(p_s s_1 + p_d(1-s_1))^2} \left(\frac{m_1}{s_1} + (1-\theta)(n - m_1/s_1)w_1 - \frac{\sum_{\ell \neq 1} m_\ell}{1-s_1} - (1-\theta) \sum_{\ell \neq 1} \frac{ns_\ell - m_\ell}{1-s_1} w_\ell \right) \end{aligned}$$

By assumption, $m_1/s_1 < \sum_{\ell \neq 1} m_\ell/(1-s_1)$ and $w_1 \leq w_\ell$, thus,

$$\begin{aligned}\frac{\partial T(w_1|p_s)}{\partial p_s} &< \frac{w_1(1-\theta)}{(p_s s_1 + p_d(1-s_1))^2}\left(n - m_1/s_1 - \sum_{\ell \neq 1} \frac{n s_\ell - m_\ell}{1-s_1}\right) \\ &= \frac{w_1(1-\theta)}{(p_s s_1 + p_d(1-s_1))^2}\left(n - m_1/s_1 - n + \frac{m - m_1}{1-s_1}\right) \\ &= \frac{w_1(1-\theta)}{(p_s s_1 + p_d(1-s_1))^2}\frac{s_1 m - m_1}{s_1(1-s_1)} < 0\end{aligned}$$

Thus, the beliefs of the agents on island 1 decrease following an increase in p_s. Then observe that for other islands $\ell \neq 1$:

$$\left(\frac{1}{1-\theta} - (1-\theta)\frac{p_s(s_\ell n - m_\ell)}{n(p_s s_\ell + p_d(1-s_\ell))}\right) w_\ell = \frac{p_s m_\ell + p_d \sum_{\ell' \neq \ell} m_{\ell'}}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)\sum_{\ell' \neq \ell} \frac{p_d(s_{\ell'} n - m_{\ell'})}{n(p_s s_\ell + p_d(1-s_\ell))} w_{\ell'}$$

when s_1 is sufficiently close to 1, the above simplifies to:

$$\frac{1}{1-\theta} w_\ell = \frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))} w_1$$

Both $\frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))}$ and $\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))}$ are decreasing in p_s, and we just showed that w_1 is decreasing in p_s. Thus belief of island ℓ is also decreasing in p_s.

2. $\underline{p_d}$: Recall that

$$\frac{\partial T(w_1|p_d)}{\partial p_d} = -\frac{p_s}{p_d}\frac{\partial T(w_1|p_s)}{\partial p_s} > 0$$

And the expression for w_ℓ when s_1 is sufficiently close to 1 is:

$$\frac{1}{1-\theta} w_\ell = \frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))} w_1$$

Both $\frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))}$ and $\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))}$ are increasing in p_d, and we showed prior that w_1 is increasing in p_d. Thus belief of island ℓ is also increasing in p_d.

3. Majorization: We consider a “Robin Hood” operation that adds a knowledgeable agent to the large island. Once again we compute the directional derivative along the gradient

$\mathbf{u} = \mathbf{e}_1 - \mathbf{e}_{\ell^*}$ for some island ℓ^*:

$$\frac{\partial T(w_1|\mathbf{m})}{\partial m_1} - \frac{\partial T(w_1|\mathbf{m})}{\partial m_{\ell^*}} = \frac{p_s(1-(1-\theta)w_1) - p_d(1-(1-\theta)w_\ell)}{p_s s_1 + (1-s_1)p_d} > 0$$

because $w_1 \leq w_\ell$ and $p_s \geq p_d$. Similarly, when s_1 is sufficiently close to 1:

$$\frac{1}{1-\theta} w_\ell = \frac{p_d m_1}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)\frac{p_d(s_1 n - m_1)}{n(p_s s_\ell + p_d(1-s_\ell))} w_1$$

which is increasing in m_1 given that w_1 is increasing in m_1.

Note that when s_1 is sufficiently large, a reverse “Robin Hood” operation that does not impact island 1 will have little effect on the beliefs of the islands. Thus, provided that no island’s centrality lies directly at $(1-b)/2$ (which holds for b on a set of full measure), this operation will have no impact on which agents are manipulated.

∎

Section 5

Illustrative Example

Suppose there are two islands of equal size. We assume that 5% of the population is knowledgeable, and we have a fixed homophily structure of $(p_s, p_d) = (0.5, 0.2)$. Moreover, suppose that $b = 0$, so that agents choose the action corresponding to the state they believe is more likely. We explore how inequality, in the form of the distribution of knowledgeable agents across the islands, affects the principal’s strategy as we vary the cost ε:

1. *Extreme inequality*: Suppose knowledgeable agents constitute 10% of the population of the first island and the second island has no knowledgeable agents at all. Let κ_ℓ denote the proportion of agents targeted by the principal on island ℓ. Then the fraction (of the total population) of agents manipulated, along with the proportions of island 1 and island 2 targeted by the principal (i.e., κ_1 and κ_2) are given in Figure B-14. With extreme inequality, the principal “gives up” on the island with more knowledgeable agents, and sends no misinformation to any agents on this island ($\kappa_1 = 0$). On the other hand, he sends misinformation to almost all of the second island and manipulates everyone on that island, until the cost of sending signals exceeds a threshold $\bar{\varepsilon} > 1$.

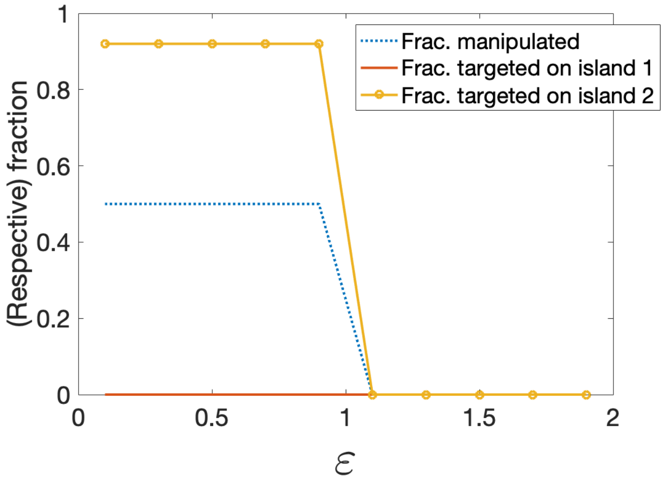


Figure B-14. The principal's optimal strategy for various values of ε under extreme inequality. Depending on the curve, (Respective) fraction refers to either the fraction of the entire population manipulated, or the fraction of a given island that is targeted. For example, when $\epsilon = 0.5$, the principal targets no one on the first island and almost everyone on the second island and ends up manipulating half the population.

2. *Intermediate inequality*: Now suppose the first island has 7.5% knowledgeable agents and the second island has only 2.5%. The principal's strategy and resulting manipulation are shown in Figure B-15. The principal sends misinformation to everyone on the second island, but importantly, this alone is not enough to manipulate the agents on that island: he also has to send misinformation to the first island in order to be able to manipulate the second island. However, this strategy targets more agents on the whole than when there is extreme inequality, and thus is more expensive. After a point $\bar{\varepsilon} < 1$, the principal has no profitable strategy. This network is therefore more resilient than one with extreme inequality, since the cost range that allows the principal to (profitably) spread misinformation is smaller.

3. *No inequality*: Finally, suppose both islands have 5% knowledgeable agents. The plot of the principal's strategy and resulting manipulation are shown in Figure B-16. Similar to the case of extreme inequality, there is a threshold $\bar{\varepsilon} > 1$ such that the principal has no profitable strategy above that threshold, and so this inequality structure is again less resilient than the case of intermediate inequality. However, unlike the case of extreme inequality, the entire population is manipulated before this threshold is met. Thus, for $\varepsilon < 1$, the absence of inequality leads to maximal manipulation relative to the other

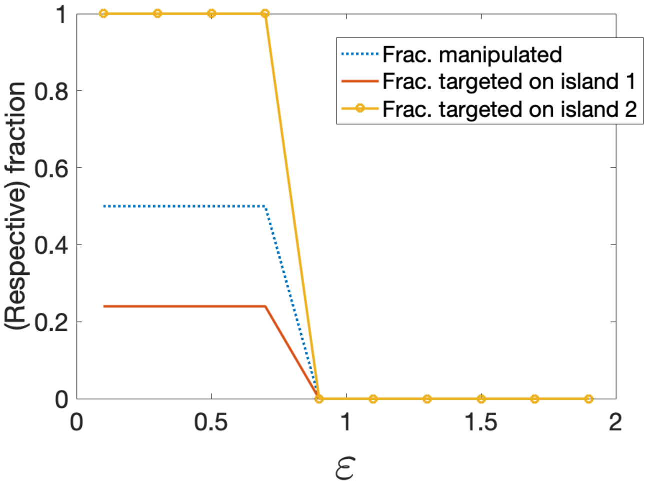


Figure B-15. The principal's optimal strategy for various values of ε under *some* inequality. Depending on the curve, (Respective) fraction refers to either the fraction of the entire population manipulated, or the fraction of a given island that is targeted.

inequality structures.

The computation of the principal's optimal strategy when there are two islands with the same population, as in the example above, can be easily generalized via the algorithm presented next.

Algorithm. Let the marginalized island be the island with fewer knowledgeable agents and the privileged island be the island with more knowledgeable agents (if the number of knowledgeable agents is equal, label them arbitrarily). Moreover, let κ_m, κ_p be the proportion of agents targeted on the marginalized and privileged islands, respectively. The principal's optimal strategy can be computed as follows:

(i) Consider the strategy where the principal targets all agents.

(a) If neither island is manipulated, the principal's optimal strategy is $\mathbf{x} = \mathbf{0}$.

(b) If both islands are manipulated, then decrease κ_m until the belief on the marginalized island matches that of the privileged island or $\kappa_m = 0$. Then decrease κ_m and κ_p one-for-one,[8] (if $\kappa_m = 0$, just decrease κ_p), until both (identical) beliefs fall below the cutoff. (Note that if $\kappa_p = \kappa_m = 0$, no island will be manipulated, so such κ_p, κ_m always exist.) Record the payoff $1 - \frac{\kappa_p + \kappa_m}{2}\varepsilon$.

[8] Formally, "one-for-one" here means decrease them simultaneously so that the beliefs of both island remain identical as these decrease, which may not necessarily correspond to the same change for κ_m as κ_p to achieve this.

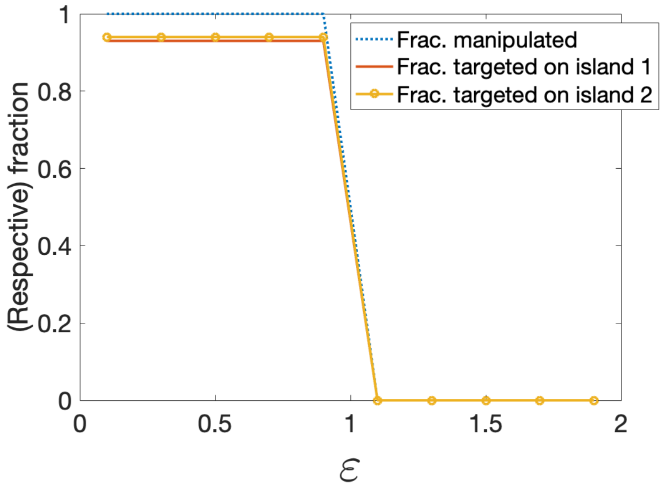


Figure B-16. The principal's optimal strategy for various values of ε under no inequality. Depending on the curve, (Respective) fraction refers to either the fraction of the entire population manipulated, or the fraction of a given island that is targeted.

(c) If just the marginalized island is manipulated, decrease κ_p until the marginalized island's belief is below π^*. If $\kappa_p = 0$ still results in the marginalized island's manipulation, begin decreasing κ_m until this island is no longer manipulated. If the payoff of $\frac{1}{2} - \frac{\kappa_p+\kappa_m}{2}\varepsilon > 0$, this is the principal's optimal strategy.

(ii) If case (i)(b) holds, consider the strategy where the principal targets no one and no island is manipulated. The principal then increases κ_m until the marginalized island is manipulated; if the marginalized island is not manipulated at $\kappa_m = 1$, then the principal begins increasing κ_p until the marginalized island is manipulated. (If $\kappa_p = \kappa_m = 1$ and the marginalized island is still not manipulated, then default to case (i)(a).) Record the payoff $\frac{1}{2} - \frac{\kappa_p+\kappa_m}{2}\varepsilon$.

(iii) Compare the recorded payoffs and the payoff of 0. If the largest payoff is 0, the principal's optimal strategy is $\mathrm{x} = \mathbf{0}$. Otherwise, the principal should employ the strategy (either (i)(b) or (ii)) corresponding to the largest payoff.

Proposition B.2.1. *The aforementioned algorithm is correct, i.e., it provides the principal's optimal strategy for any* ε.

Proof of Proposition B.2.1. The principal first decides the cheapest way to manipulate 0, 1, and 2 islands, respectively, and then compares the payoffs and chooses whichever is maximal. The cheapest way to manipulate no islands is $\mathbf{x} = \mathbf{0}$, which is always the recommended outcome of the algorithm for 0 islands.

To manipulate one island, it is always cheaper to manipulate the marginalized island. Note w_m^K does not depend on κ_m or κ_p, and that $\partial w_m^D / \partial \kappa_m > \partial w_m^D / \partial \kappa_p$. Thus, the optimal way to manipulate the marginalized island is to decrease κ_p as much as possible, leaving $\kappa_m = 1$, until the marginalized island is no longer manipulated. If it is still manipulated at $\kappa_p = 0$, then the only more cost effective strategy would be to decrease κ_m until you lose this island. This is precisely the strategy identified in (i)(c) and (ii). In the case of (i)(c), we know this to be the optimal strategy if it beats manipulating no one, because it is impossible to manipulate both islands.

To manipulate two islands, we show that it is always optimal for either: (1) $\kappa_m = 0$ and κ_p is chosen smallest to manipulate both islands, or (2) both islands have the same belief at the optimal strategy (right at the cutoff $(1+b)/2$). Recall that $\partial w_\ell^D / \partial \kappa_\ell > \partial w_\ell^D / \partial \kappa_{\ell'}$ for both islands, where $\ell' \neq \ell$. The privileged island will have beliefs closer to the truth when $\kappa_m = \kappa_p$, it must be the case that $\kappa_p \geq \kappa_m$ in the optimal strategy for 2-island manipulation. Of these strategies, (1) is clearly then optimal provided that $\kappa_m = 0$ does not cause the marginalized island's belief to lie above the cutoff (and thus, not manipulate both islands). Otherwise, if some island ℓ's belief is strictly above the cutoff, then one can decrease κ_ℓ a small amount and increase $\kappa_{\ell'}$ for $\ell' \neq \ell$ by a smaller amount, again, because $\partial w_\ell^D / \partial \kappa_\ell > \partial w_\ell^D / \partial \kappa_{\ell'}$ for both islands ℓ. This is cheaper than the previous strategy, a contradiction. So (2) must be optimal.

Finally, step (iii) checks whether 0-island, 1-island, or 2-island manipulation is the most profitable. ∎

Proof of Theorem 3.2.3. The beliefs of all agents will be identical in the network with the least inequality. Let $m(b)$ be the maximum number of knowledgeable agents such that if these agents are all distributed evenly across the islands, every agent is manipulated if the principal targets every DeGroot, which is a function of b. Moreover, because of symmetry, it is clear that manipulating every agent or manipulating no agent is the principal's optimal strategy, and in particular for $\varepsilon < 1$, manipulating every agent is a profitable strategy. (Note that this does not imply that $\boldsymbol{\gamma} = \mathbf{1}_D$ is optimal, just that manipulating every agent is optimal.) When we move to an inequality configuration with the most inequality, there are two cases: (1) $m(b) \leq n/k - 1$ or (2) $m(b) \geq n/k$. In case (1), we stack all of the knowledgeable agents on a single island and set $p_s = 1$ and $p_d = 0$, noting that there is at least one DeGroot on this island. It is clear that this means the island with all the knowledgeable agents will have a decrease in their DeGroot centrality, which by definition of $m(b)$, will protect this island from manipulation. At the same time, this configuration is still susceptible to manipulation, because the islands with all DeGroots agents will be manipulated given $\varepsilon < 1$. In case (2), we make island 1 contain all knowledgeable agents and one DeGroot, and then distribute the remaining knowledgeable agents equally amongst the rest of the islands. For sufficiently large n, this will always (strictly) decrease the centrality of the one DeGroot agents on the island with concentrated knowledgeable agents, thereby protecting her from manipulation; at the same time, the DeGroots on the other islands will continue to be manipulated, as their centrality does not decrease (and may increase).

Finally, we show there exists a model with intermediate inequality that is impervious for some open interval of $\varepsilon \in (\varepsilon^*, \varepsilon^{**})$ and $b \in (b^*, b^{**})$. Once again there are two cases: (1) $m(b) \leq n/k - 1$ and (2) $m(b) \geq n/k$. In the former case, put all of the knowledgeable agents on island 1 along with one DeGroot, as before. In the latter case, put $n/k - 1$ knowledgeable agents on island 1 along with one DeGroot, and then distribute the rest of the knowledgeable agents evenly amongst the remaining islands. If $0 < p_d < p_s < 1$, then we have a model of intermediate inequality where the beliefs (of the correct state) of the agents on island 1 exceed those on the other islands (which are identical because of symmetry).[9] Because the beliefs of the agents on island 1 exceed that of other islands, we know the principal cannot manipulate the DeGroot on island 1, even if he were to target every DeGroot in the population. Next, we show that for any $\delta < 1$, there exists some p_d and $b > -1$ such that the principal needs to target at least δ proportion of the DeGroots on island 1 *and* at least δ proportion of the DeGroots on the rest of the islands.[10] We know that $\pi_1 > \pi_\ell$ when the principal targets every DeGroot for all $\ell \neq 1$ given that $p_d < p_s$. Thus, we can always choose b such that $(1+b)/2$ is arbitrarily close to π_ℓ but still satisfies $\pi_1 > (1+b)/2 > \pi_\ell$. Given $p_d > 0$, any (substantial) deviation from $\boldsymbol{\gamma} = \mathbf{1}_D$, in the ∞-norm, leads to some island $\ell \neq 1$ not being manipulated. Because the principal should always enact a symmetric strategy with respect to all islands $\ell \neq 1$, we then either have

[9] This was shown in the proof of Theorem 3.2.1.

[10] Note that there is technically only one DeGroot on island 1, so targeting δ proportion of one DeGroot seems non-sensical. However, Appendix B.2.1 reconciles this: because the analysis is always closed under multiplication of each island by the same proportion of knowledgeable agents and DeGroots, we can always expand the size of the DeGroot population such that δ proportion (assuming $\delta \in \mathcal{Q}$) of $\delta(n/k - m_1)$ is an integer.

(i) the network is impervious, or (ii) the principal should send signals to at least δ proportion of each island, where δ can be arbitrarily close to 1. In the latter case, the payoff of the principal is no more than $n\left(\frac{k-1}{k}\right) - m(b) - (n - m(b))\delta\varepsilon$, again, for δ arbitrarily close to 1. Thus, there exists $\varepsilon < 1$ such that the principal's payoff from this strategy is negative. Hence, the network is impervious with a model of intermediate inequality. ∎

Section 6

Proof of Proposition 3.2.1. Suppose there are n agents in the network, and there are m knowledgeable agents. We denote by $w_{\ell\to r}$ the weighted walks from an agent on island ℓ to any knowledgeable agent on island r. For $\ell \neq r$, we can write:

$$\begin{aligned}
w_{\ell\to r} &\geq (1-\theta)\frac{p_d m_r}{n(p_s s_\ell + np_d(1-s_\ell))} + (1-\theta)^2 \left(\frac{np_s s_\ell w_{\ell\to r} + p_d(ns_r - m_r)w_{r\to r} + np_d \sum_{\tau\neq r,\ell} s_\tau w_{\tau\to r}}{np_s s_\ell + np_d(1-s_\ell)}\right) \\
&\geq (1-\theta)\frac{p_d m_r}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)^2 \Bigg(\frac{np_s s_\ell \underline{w}_r + p_d(ns_r - m_r)(\underline{w}_r + w_{r\to r} - \underline{w}_r) + np_d(1 - s_\ell - s_r)\underline{u}}{np_s s_\ell + np_d(1-s_\ell)} \\
&= (1-\theta)\frac{p_d m_r}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)^2 \left(\underline{w}_r + \frac{np_d s_r(w_{r\to r} - \underline{w}_r) - p_d m_r w_{r\to r}}{np_s s_\ell + np_d(1-s_\ell)}\right) \\
&= (1-\theta)\frac{p_d m_r}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)^2 \left(\frac{p_s s_\ell + p_d(1-s_\ell-s_r)}{p_s s_\ell + p_d(1-s_\ell)}\underline{w}_r + \frac{p_d(ns_r - m_r)}{n(p_s s_\ell + p_d(1-s_\ell))}w_{r\to r}\right) \\
&\geq (1-\theta)^2\frac{p_d m_r}{n(p_s s_\ell + p_d(1-s_\ell))} + (1-\theta)^2 \left(\frac{p_s s_\ell + p_d(1-s_\ell-s_r)}{p_s s_\ell + p_d(1-s_\ell)}\underline{w}_r + \frac{p_d(ns_r - m_r)}{n(p_s s_\ell + p_d(1-s_\ell))}w_{r\to r}\right)
\end{aligned}$$

where $\underline{w}_r = \min_{\tau\neq r} w_{\tau\to r}$. This implies that:

$$\underline{w}_r \geq (1-\theta)^2 \frac{p_d m_r + p_d(ns_r - m_r)w_{r\to r}}{n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r}$$

Similarly,

$$\begin{aligned}
w_{r\to r} &= (1-\theta)\frac{p_s m_r}{n(p_s s_r + p_d(1-s_r))} + (1-\theta)^2 \frac{p_s(ns_r - m_r)w_r^r + p_d \sum_{\tau\neq r} s_\tau w_r^\tau}{np_s s_r + np_d(1-s_r)} \\
&\geq (1-\theta)\frac{p_s m_r}{n(p_s s_r + p_d(1-s_r))} + (1-\theta)^2 \left[\underline{w}_r + \frac{np_s s_r(w_{r\to r} - \underline{w}_r) - p_s m_r w_{r\to r}}{np_s s_r + np_d(1-s_r)}\right] \\
&= (1-\theta)\frac{p_s m_r}{n(p_s s_r + p_d(1-s_r))} + (1-\theta)^2 \left[\frac{p_d(1-s_r)}{p_s s_r + p_d(1-s_r)}\underline{w}_r + \frac{p_s(ns_r - m_r)}{n(p_s s_r + p_d(1-s_r))}w_{r\to r}\right] \\
&\geq (1-\theta)^2\frac{p_s m_r}{n(p_s s_r + p_d(1-s_r))} + (1-\theta)^2 \left[\frac{p_d(1-s_r)}{p_s s_r + p_d(1-s_r)}\underline{w}_r + \frac{p_s(ns_r - m_r)}{n(p_s s_r + p_d(1-s_r))}w_{r\to r}\right]
\end{aligned}$$

which moreover implies that

$$w_{r\to r} \geq (1-\theta)\frac{p_s m_r + (1-\theta)np_d(1-s_r)\underline{w}_r}{\theta p_s n s_r + np_d(1-s_r) + (1-\theta)p_s m_r}$$

Combining these two results we get:

$$\begin{aligned}
&\underline{w}_r \geq (1-\theta)^2\frac{p_d m_r + p_d(ns_r - m_r)(1-\theta)\frac{p_s m_r + (1-\theta)np_d(1-s_r)\underline{w}_r}{\theta p_s n s_r + np_d(1-s_r)+(1-\theta)p_s m_r}}{n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r} \\
\implies &\left[n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r\right]\underline{w}_r \\
&\geq (1-\theta)^3\frac{p_d m_r + p_d(ns_r - m_r)\frac{p_s m_r + (1-\theta)np_d(1-s_r)\underline{w}_r}{\theta p_s n s_r + np_d(1-s_r)+(1-\theta)p_s m_r}}{n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r} \\
\implies &\left[n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r\right]\underline{w}_r \\
&\geq (1-\theta)^3 p_d m_r + (1-\theta)^3 p_d(ns_r - m_r)\frac{p_s m_r}{\theta p_s n s_r + np_d(1-s_r) + (1-\theta)p_s m_r} \\
&+ (1-\theta)^4 p_d(ns_r - m_r)\frac{np_d(1-s_r)}{\theta p_s n s_r + np_d(1-s_r) + (1-\theta)p_s m_r}\underline{w}_r
\end{aligned}$$

Note that:

$$\begin{aligned}
&\Big(\left[n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r\right]\left[\theta p_s n s_r + np_d(1-s_r) + (1-\theta)p_s m_r\right] - (1-\theta)^4 p_d^2(ns_r - m_r)(1-s_r)\Big) \\
\geq&(1-\theta)^3 p_d m_r\left[\theta p_s n s_r + np_d(1-s_r) + (1-\theta)p_s m_r\right] + (1-\theta)^3 p_d(ns_r - m_r)
\end{aligned}$$

Therefore, we can write $\underline{w}_r \geq N(n)/D(n)$, where:

$$\begin{aligned}
N(n) \equiv&(1-\theta)^3 p_d m_r\left[\theta p_s n s_r + np_d(1-s_r) + (1-\theta)p_s m_r\right] + (1-\theta)^3 p_d(ns_r - m_r) \\
D(n) \equiv&\left[n\theta(p_s s_\ell + p_d(1-s_\ell)) - (1-\theta)p_d s_r\right]\left[\theta p_s n s_r + np_d(1-s_r) + (1-\theta)p_s m_r\right] - (1-\theta)^4 p_d^2(ns_r - m_r)(1 -
\end{aligned}$$

If there are $m = c_r n$ knowledgeable agents on island r, as $n \to \infty$, we have that:

$$\underline{w}_r \geq \frac{(1-\theta)^3 p_d c_r}{\theta(p_s\bar{s} + p_d(1-\bar{s}))}$$

where $\bar{s} = \max_{\tau\in\{1,\ldots,k\}} s_\tau$. Thus, the network is impervious as long as $\underline{w}_r > (1+b)/2$. This moreover implies the network is impervious if $c_r \geq \frac{\theta(p_s\bar{s}+p_d(1-\bar{s}))(1+b)}{2(1-\theta)^3 p_d}$ for any island r. By the pigeonhole principle, there must be an island with at least c/k proportion of the population that is knowledgeable. Thus taking $c = k\frac{\theta(p_s\bar{s}+p_d(1-\bar{s}))(1+b)}{2(1-\theta)^3 p_d}$ and applying the result for the island

r which has the largest proportion of knowledgeable agents, we see the network is impervious to manipulation. Moreover $c < 1$ provided that θ is not too large. ∎

Proof of Proposition 3.2.2. For each θ and c we construct a strong inequality model where all but $\bar{k}$ communities are manipulated. Put all cn knowledgeable agents on the first island on the line topology and let all other islands contain only DeGroots and be the same size as each other. We assume that the principal attempts to manipulate the last $k - \bar{k}$ communities along the line. We compute $\mathcal{D}_\ell(\mathbf{1})$ for every island by counting knowledgeable walks for every island; we denote these walks by w_ℓ which is equivalent to $1 - \mathcal{D}_\ell(\gamma)$. For island 2, we have the recursion:

$$\begin{aligned} w_2 &= (1-\theta)\frac{p_d s_1}{p_d(s_1+s_3)+p_s s_2} + (1-\theta)^2 \frac{p_s s_2 w_2 + p_d s_3 w_3}{p_d(s_1+s_3)+p_s s_2} \\ \implies w_2 &= (1-\theta)\frac{p_d s_1}{p_d(s_1+s_3)+p_s s_2 - (1-\theta)^2 p_s s_2} + (1-\theta)^2 \frac{p_d s_3 w_3}{p_d(s_1+s_3)+p_s s_2 - (1-\theta)^2 p_s s_2} \end{aligned}$$

For $\ell \geq 3$:

$$\begin{aligned} w_\ell &= (1-\theta)^2 \frac{p_d(w_{\ell-1}s_{\ell-1} + w_{\ell+1}s_{\ell+1}) + p_s w_\ell s_\ell}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell} \\ \implies w_\ell &= (1-\theta)^2 \frac{p_d(w_{\ell-1}s_{\ell-1} + w_{\ell+1}s_{\ell+1})}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell} \end{aligned}$$

with $w_1 = 1$ because the island consists of all knowledgeable agents. We first show that $\lim_{\ell\to\infty} w_\ell$ must exist. To do this, we show that w_ℓ is monotonically decreasing in ℓ. We know there is a unique fixed point for w, so if we prove that a decreasing sequence of w_ℓ maps to another decreasing sequence of w_ℓ, then by Brouwer's fixed point theorem the unique solution must be a decreasing in ℓ. Note that:

$$\begin{aligned} & w_\ell \leq (1-\theta)^2 \frac{p_d(w_{\ell-1}s_{\ell-1} + w_\ell s_{\ell+1})}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell} \\ \implies & \left(1 - \frac{p_d s_{\ell+1}}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell}\right) w_\ell \leq (1-\theta)^2 \frac{p_d w_{\ell-1} s_{\ell-1}}{p_d(s_{\ell-1}+s_{\ell+1}) + p_s s_\ell - (1-\theta)^2 p_s s_\ell} \\ \implies & w_\ell \leq (1-\theta)^2 \frac{p_d w_{\ell-1} s_{\ell-1}}{p_d s_{\ell-1} + p_s s_\ell - (1-\theta)^2 p_s s_\ell} \leq w_{\ell-1} \end{aligned}$$

where the final inequality follows from the fact that $\beta\frac{\alpha}{\alpha+\delta} < 1$ for $\alpha, \beta, \delta \in (0,1)$. Thus, w_ℓ

converges to some w_∞. Note that w_∞ must satisfy the fixed-point equation:

$$w_\infty = (1-\theta)^2 \frac{p_d(s_{\ell-1}+s_{\ell+1})w_\infty}{p_d(s_{\ell-1}+s_{\ell+1})+p_s s_\ell-(1-\theta)^2 p_s s_\ell}$$

Again, since $\frac{p_d(s_{\ell-1}+s_{\ell+1})}{p_d(s_{\ell-1}+s_{\ell+1})+p_s s_\ell-(1-\theta)^2 p_s s_\ell} < 1$, clearly $w_\infty = 0$.

Now, suppose the principal attempts to manipulate $k-\bar{k}$ communities at the end of the line. By our previous result, we know that for every $\delta > 0$, there exists a sufficiently large $\bar{k}$, such that $w_{\bar{k}+1} < \delta$. Therefore, for any b, the principal can manipulate $k-\bar{k}$ of the islands at the end of the line. This yields a payoff of $\sum_{\ell=\bar{k}+1}^{k} ns_k - n\varepsilon(1-s_1)$, which is positive for some sufficiently small $\varepsilon > 0$. Thus, the network is susceptible to manipulation and all but $\bar{k}$ islands are manipulated. ∎

Section 7

Proof of Corollary 3.2.1. By assumption, the budget is large enough to make the network impervious. Thus, by Theorem 3.2.1, the network is impervious if the current distribution of knowledgeable agents $\mathbf{m}$ is majorized by every other distribuiton (i.e., inequality cannot be reduced by a more equal redistribution of knowledgeable agents). Minimizing inequality with an educational intervention accomplishes this. ∎

Proof of Corollary 3.2.2. Leveraging Theorem 3.2.4, we know that decreasing inequality when the big island is the most underprivileged does not introduce more manipulation, and in fact, might reduce it. Assuming the policy does not put all of the knowledgeable agents on island 1, we know such a redistribution is a different feasible policy. In the proof of Theorem 3.2.4, this is shown to be true because decreasing inequality increases all agents' beliefs. This is exactly the definition of a dominant policy. ∎

Proof of Proposition 3.2.3. We show that if the knowledgeable agents are assigned in proportion to island populations, i.e., $m_\ell = M \cdot s_\ell$, then all DeGroot centralities are equal. Then, homophily has no effect (beliefs are the same on every island), so setting $p_d = p_d^o$ is optimal, and always feasible because it costs nothing. Consider the map T:

$$T\;:\;\mathbf{w}\mapsto(1{-}\theta)\begin{pmatrix}\frac{p_s m_1+\sum_{\ell\neq 1}p_d m_\ell}{np_s s_1+np_d(1-s_1)}\\ \cdots\\ \frac{p_s m_k+\sum_{\ell\neq k}p_d m_\ell}{np_s s_k+np_d(1-s_k)}\end{pmatrix}+(1{-}\theta)^2\begin{pmatrix}\frac{p_s(ns_1-m_1)}{np_s s_1+np_d(1-s_1)} & \frac{p_d(ns_2-m_2)}{np_s s_1+np_d(1-s_1)} & \cdots & \frac{p_d(ns_k-m_k)}{np_s s_1+np_d(1-s_1)}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{p_d(ns_1-m_1)}{np_s s_k+np_d(1-s_k)} & \frac{p_d(ns_2-m_2)}{np_s s_k+np_d(1-s_k)} & \cdots & \frac{p_s(ns_k-m_k)}{np_s s_k+np_d(1-s_k)}\end{pmatrix}\mathbf{w}$$

Note we will simply plug in $\mathbf{w}=w^*\mathbf{1}$ to T and show it is a fixed point for some constant w^*. For island ℓ:

$$\begin{aligned}w_\ell &= (1-\theta)\cdot\frac{p_s m_\ell+\sum_{\ell'\neq\ell}p_d m_{\ell'}+(1-\theta)p_s(ns_\ell-m_\ell)w_\ell+(1-\theta)\sum_{\ell'\neq\ell}p_d(ns_{\ell'}-m_{\ell'})w_{\ell'}}{np_s s_\ell+np_d(1-s_\ell)}\\ &= (1-\theta)\cdot\frac{p_s s_\ell M+\sum_{\ell'\neq\ell}p_d s_{\ell'}M+(1-\theta)p_s(ns_\ell-s_\ell M)w^*+(1-\theta)\sum_{\ell'\neq\ell}p_d(ns_{\ell'}-s_{\ell'}M)w^*}{np_s s_\ell+np_d(1-s_\ell)}\\ &= (1-\theta)\cdot\frac{M(p_s s_\ell+p_d(1-s_\ell))+(1-\theta)w^*(np_s s_\ell+np_d(1-s_\ell)-M(p_s s_\ell+p_d(1-s_\ell)))}{np_s s_\ell+np_d(1-s_\ell)}\\ &= (1-\theta)\cdot\frac{M+(1-\theta)w^*(n-M)}{n}\end{aligned}$$

The above expression has no dependence on ℓ. Letting $w^*=\frac{M(1-\theta)}{M(1-\theta)+\theta n}$, we see then that $w_\ell=w^*$, which completes the proof. This has no dependence on p_d, so all p_d are optimal, including p_d^o. ∎

Proof of Proposition 3.2.4. The condition that the budget exceeds $\phi(p_s-p_d^o)$ is to guarantee that $p_d=p_s$ is feasible. By Theorem 3.2.1, given that all islands are the same size, a network with less inequality cannot transition from impervious to susceptible. Thus, removing all homophily (i.e., reducing inequality the most through the homophily parameters) must make the network impervious given this is possible for some homomphily structure. Thus, setting $p_s=p_d$ makes the network impervious which is obviously an optimal policy. ∎

B.3 Contrasting Bayesian and DeGroot Models

B.3.1 Proofs

We first provide two auxillary lemmas that we use in the proofs of our main results.

Auxiliary Lemmas

Recall that H is the distribution of prior beliefs in the population and p is the strength of organic news (likelihood a message corresponds to the true state θ). For misinformation, q denotes the amount of misinformation in the system, r denotes how much of this misinformation argues for state R, and F is the distribution of r.

Lemma B.3.1. *Let H be symmetric. If $\theta = L$, the DeGroot society mislearns if and only if $r \geq (1-2(1-q)(1-p))/(2q)$; if $\theta = R$, the DeGroot society mislearns if and only if $r \leq (1-2(1-q)p)/(2q)$.*

Proof of Lemma B.3.1. We prove this for $\theta = L$; the case of $\theta = R$ is similar. For a fixed realization $r \sim F(\cdot)$, every DeGroot agent i converges to belief π_∞ about $\theta = R$:

$$\pi_\infty = \int_0^1 \left(((1-p)(1-q)+qr)\frac{p\alpha}{p\alpha+(1-p)(1-\alpha)} + (p(1-q)+q(1-r))\frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)} \right) h(\alpha)\, d\alpha$$

Via the Leibniz integral rule, we see that:

$$\begin{aligned}
\frac{d\pi_\infty}{dr} &= \int_0^1 \frac{\partial}{\partial r} \left((1-p+qr)\frac{p\alpha}{p\alpha+(1-p)(1-\alpha)} + (p+q(1-r))\frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)} \right) h(\alpha)\, d\alpha \\
&= q\int_0^1 \left(\frac{p\alpha}{p\alpha+(1-p)(1-\alpha)} - \frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)} \right) h(\alpha)\, d(\alpha) \\
&= q\int_0^1 \frac{(1-\alpha)\alpha(2p-1)}{(p-\alpha(2p-1))((1-p)+\alpha(2p-1))} h(\alpha)\, d\alpha
\end{aligned}$$

Note that all expressions are positive because $p > 1/2$. The only non-trivial one to verify is the first expression in the denominator, which is linear in α and thus it is sufficient to verify it is non-negative for all $\alpha \in \{0,1\}$ to prove it is non-negative for all $\alpha \in [0,1]$. When $\alpha = 0$ it is equivalent to p and when $\alpha = 1$ it is equivalent to $1-p$.

Thus, $d\pi_\infty/dr > 0$ for all r. Consider the expression for $\pi_\infty(r)$ when $r = \tilde{r} \equiv (1-2(1-q)(1-p))/(2q)$:

$$\begin{aligned}
&\int_0^1 \left(((1-p)(1-q)+q\tilde{r})\frac{p\alpha}{p\alpha+(1-p)(1-\alpha)} + (p(1-q)+q(1-\tilde{r}))\frac{(1-p)\alpha}{(1-p)\alpha+p(1-\alpha)} \right) h(\alpha)\, d\alpha \\
&= \int_0^{1/2} \frac{p\alpha}{2(p\alpha+(1-p)(1-\alpha))} h(\alpha)\, d\alpha + \int_{1/2}^1 \frac{(1-p)\alpha}{2((1-p)\alpha+p(1-\alpha))} h(\alpha)\, d\alpha
\end{aligned}$$

For the second integral expression, we make the change of variables $\beta \equiv 1 - \alpha$, which yields:

$$= \int_0^{1/2} \frac{p\alpha}{2(p\alpha + (1-p)(1-\alpha))} h(\alpha)\, d\alpha + \int_0^{1/2} \frac{(1-p)(1-\beta)}{2((1-p)(1-\beta) + p\beta)} h(1-\beta)\, d\beta$$

By symmetry of H, we know that $h(1-\beta) = h(\beta)$, thus the above expression simplifies to $\int_0^{1/2} h(\alpha)\, d\alpha = 1/2$ because H is symmetric and $\int_0^1 h(\alpha)\, d\alpha = 1$. Similarly, because $d\pi_\infty/dr > 0$, we know whenever $r > \tilde{r}$, $\pi_\infty(r) > 1/2$ and so agents elect action $a_i = R$, and the society mislearns. Whenever $r < \tilde{r}$, $\pi_\infty(r) < 1/2$ and so agents elect action $a_i = L$, and the society *does* learn. ∎

Lemma B.3.2. *If H has full support and $\theta = L$, the Bayesian society mislearns if and only if $r \geq \underline{r} + \frac{(2p-1)(1-q)}{q}$; if $\theta = R$, the Bayesian society mislearns if and only if $r \leq \bar{r} - \frac{(2p-1)(1-q)}{q}$.*

Proof of Lemma B.3.2. Once again, we prove this $\theta = L$ and remark that the case of $\theta = R$ is similar. Unlike Lemma B.3.1, we prove both the “if” and “only if” parts separately:

(i) *If part*: Suppose $r \geq \underline{r} + \frac{(2p-1)(1-q)}{q}$. There is some proportion ρ_L of messages that advocate for L and some proportion $\rho_R = 1 - \rho_L$ of messages that advocate for R.

When $\theta = L$, $p(1-q)$ proportion of the messages are both organic and advocate for L, i.e., $m_i = L$, when the population is large. Similarly, $q(1-r)$ proportion of messages are inorganic (“misinformation”) and advocate for L as well. Call this scenario 1.

When $\theta = R$, $(1-p)(1-q)$ proportion of the messages are both organic and advocate for L, with once again, $q(1-r)$ proportion of messages that are inorganic and advocate for L too. Call this scenario 2.

Both scenarios occur with positive probability if there exist some r_1 and r_2, both within $[\underline{r}, \bar{r}]$, where both scenarios yield the same realized distribution ρ_L and ρ_R. In scenario 1 we admit $\rho_L^1 = p(1-q) + q(1-r_1)$ and in scenario 2 we admit $\rho_L^2 = (1-p)(1-q) + q(1-r_2)$. Because $\theta = L$, we know that $r_1 = r$ and scenario 1 occurs with positive probability. To see if scenario 2 occurs with positive probability, one needs to find the existence of $r_2 \in [0,1]$ such that $p(1-q) + q(1-r) = (1-p)(1-q) + q(1-r_2)$. When $r_2 = 1$ and since $p(1-q) + q(1-r) > (1-p)(1-q)$, there exists a value for r_2 where the left-hand side is greater than the right-hand side. Moreover, the left-hand side is decreasing in

r_2, so verify there exists some r_2 where equality can be obtained, it is sufficient to have $p(1-q)+q(1-r) \leq (1-p)(1-q)+q$. Rearranging gives the condition in the lemma.

Finally, we note that under this condition, both scenario 1 and scenario 2 occur with probability $\eta_1, \eta_2 > 0$. The probability that an agent with prior $\pi_{i,0}$ about scenario 2 (i.e., $\theta = R$) is:

$$\pi_{i,2} = \frac{\eta_2 \pi_{i,0}}{\eta_2 \pi_{i,0} + \eta_1(1-\pi_{i,0})}$$

Taking $\pi_{i,0}$ sufficiently close to 1 yields $\pi_{i,2} > 1/2$, and given H has full support, implies some positive fraction of the Bayesian population mislearns, so society fails to learn as well.

(ii) *Only if part*: Suppose $r < \underline{r} + \frac{(2p-1)(1-q)}{q}$. Then by the same argument in the "if" proof, there exists no value for r_2 such that $p(1-q)+q(1-r) = (1-p)(1-q)+q(1-r_2)$ because $p(1-q)+q(1-r) > (1-p)(1-q)+q(1-r_2)$ for all $r_2 \in [\underline{r}, \bar{r}]$. Thus, there exists a unique value for r that yields message distribution ρ_L (scenario 1) and it necessarily corresponds to $\theta = L$. Note that:

$$\pi_{i,2} = \frac{\eta_2 \pi_{i,0}}{\eta_2 \pi_{i,0} + \eta_1(1-\pi_{i,0})} = 0$$

given that $\eta_1 > 0$ and $\eta_2 = 0$, and $\pi_{i,0} \in (0,1)$ almost surely. Thus, all of the Bayesian agents learn the correct state $\theta = L$. ∎

Proofs of Section 4

Proof of Proposition 1. We showed in Lemma B.3.1 that $d\pi_\infty/dr > 0$ without utilizing the symmetry assumption on H. The DeGroot society mislearns if and only if that $\pi_\infty(r) > 1/2$ when $\theta = L$. Thus, there is a unique cutoff r_D^* such that the DeGroot society mislearns if and only if $r > r_D^*$. ∎

Proof of Proposition 2. We showed in Lemma B.3.2 there is a single narrative when $r < r_B^* \equiv \underline{r} + \frac{(2p-1)(1-q)}{q}$, which implies the Bayesian society must learn in this setting, as $\pi_{i,2} = 0$ when $\theta = L$. ∎

Proof of Theorem 1. Note by Lemma B.3.1, the DeGroot society mislearns with positive probability when $q > \frac{2p-1}{2p}$. By the arguments in Lemma B.3.2, given that H may or may not have full support, a necessary (but not necessarily sufficient) condition for the Bayesian society to mislearn is that $q > \frac{2p-1}{2p-\underline{r}} \geq \frac{2p-1}{2p}$. This establishes part (a) by taking $q^* = \frac{2p-1}{2p}$.

For part (b), when $q > q^*$ and H and F have full support (so $\underline{r} = 0$), the Bayesian society mislearns with probability $1-F\left(\frac{(2p-1)(1-q)}{q}\right)$ and by Lemma B.3.1 the DeGroot society mislearns with probability $1-F\left(\frac{1-2(1-q)(1-p))}{2q}\right)$. Observe that:

$$\frac{(2p-1)(1-q)}{q} - \frac{1-2(1-q)(1-p))}{2q} = \frac{2p(1-q)-1}{2q}$$

which is a decreasing function in q and is exactly equal to 0 when $q = q^*$. Thus, by monotonicity of F, $F\left(\frac{(2p-1)(1-q)}{q}\right) < F\left(\frac{1-2(1-q)(1-p))}{2q}\right)$ and so $1-F\left(\frac{(2p-1)(1-q)}{q}\right) > 1-F\left(\frac{1-2(1-q)(1-p))}{2q}\right)$, implying the Bayesian society mislearns more often. ∎

Proof of Theorem 2. When $\theta = L$, note the ratio of DeGroot mislearning to the ratio of Bayesian mislearning is given by:

$$\mu = \frac{1-F\left(\frac{1-2(1-q)(1-p)}{2q}\right)}{1-F\left(\frac{(2p-1)(1-q)}{q}\right)}$$

by Lemma B.3.1 and Lemma B.3.2. Differentiating with respect to p, we get that

$$\frac{\partial \mu}{\partial p} = \frac{f\left(\frac{(2p-1)(1-q)}{q}\right)\frac{2(1-q)}{q}\left(1-F\left(\frac{1-2(1-q)(1-p)}{2q}\right)\right) - f\left(\frac{1-2(1-q)(1-p)}{2q}\right)\frac{1-q}{q}\left(1-F\left(\frac{(2p-1)(1-q)}{q}\right)\right)}{\left(1-F\left(\frac{(2p-1)(1-q)}{q}\right)\right)^2}$$

Note that $\partial\mu/\partial p > 0$ if and only if

$$2f\left(\frac{(2p-1)(1-q)}{q}\right)\left(1-F\left(\frac{1-2(1-q)(1-p)}{2q}\right)\right) > f\left(\frac{1-2(1-q)(1-p)}{2q}\right)\Bigl(1-F\left(\frac{(2p-1)(1-q)}{q}\right)$$

It is easy to see that $q^* = \frac{2p-1}{2p}$ from the proof of Theorem 1, and thus $\alpha \equiv \frac{1-2(1-p)(1-q)}{2q}$ and $\alpha - \beta = \frac{(2p-1)(1-q)}{q}$ (given that $\beta \equiv p\left(1-\frac{q^*}{q}\right)$). Substituting we have that $\partial\mu/\partial p > 0$ if and only if

$$2f(\alpha-\beta)(1-F(\alpha)) > f(\alpha)(1-F(\alpha-\beta))$$

or in other words, $2\lambda_F(\alpha-\beta) > \lambda_F(\alpha)$, which proves the claim. ∎

Proofs of Section 5

Proof of Theorem 3. Observe that larger values of γ decrease the lower support of the distribution H: if $\bar{\pi} > 1/2$ is the upper support for h, then $\bar{\pi}_\gamma = \bar{\pi} + (\bar{\pi} - 1/2)\gamma$ is increasing in γ and when $\gamma = 0$, $\bar{\pi}_\gamma = \bar{\pi}$. Moreover, all values of γ preserve the symmetry of H.

By Lemma B.3.1, and since H is always symmetric, the probability of DeGroot mislearning does not depend of γ. For Bayesian learning, there is always either one $(r < r^*_B)$ or two narratives $(r > r^*_B)$. In the former case, Bayesians always learn. In the latter case, let η_L be the likelihood of the $\theta = L$ narrative and η_R be the likelihood of the $\theta = R$ narrative. Then the Bayesian society mislearns if and only if:

$$\frac{\eta_R \bar{\pi}}{\eta_R \bar{\pi} + \eta_L (1 - \bar{\pi})} > 1/2$$

But note that the left-hand side is increase $\bar{\pi}$, so for every realization of r, mislearning can only become “more likely” as $\bar{\pi}$ increases.[11] Integrating over all of r shows that the likelihood of Bayesian mislearning is increasing in $\bar{\pi}$ and, in particular, is increasing in γ.

Next observe that when $\gamma = -1$, the Bayesian society mislearns with lower probability than the DeGroot society. To show this, note that when $\gamma = -1$, the density h is a Dirac-delta function at belief $\pi = 1/2$. Thus, all Bayesian agents initially agree, so there is a homogenous prior. By the improvement principle (see, for instance, Golub and Sadler (2017)), Bayesian agents must be able to outperform the DeGroot heuristic.

Finally, when $\gamma = \bar{\gamma}$, then H has full support on $[0, 1]$, so by Theorem 1(b), we know the Bayesian society mislearns with higher probability than the DeGroots. By the previous paragraph, we know that when $\gamma = -1$, then the Bayesian society mislearns. Because the mislearning probability is increasing in γ for Bayesians but constant for DeGroots, there must be a unique single-crossing γ^* that determines the phase transition. ∎

Proof of Proposition 3. We prove the two parts of the result, which separate into DeGroot and Bayesian societies. For both, we fix $\theta = L$ for concreteness.

(i) *DeGroot society*: Because we have fixed $\theta = L$, we know that $m^p = L$ for targeted agents.

[11] “More likely” is a slight abuse of terminology, because for a given realization of r, the society either learns or does not almost surely. Formally, we mean that an increase in $\bar{\pi}$ can not transition society from mislearning to learning for this value of r.

Thus, there are two cases for $\pi_{i,1}$ for the DeGroot agent that depend on whether (i) agent i receives $m_i = R$ and $m^p = L$ or (ii) agent i receives both $m_i = L$ and $m^p = L$. The former case occurs with probability $(1-p)(1-q) + qr$ whereas the latter occurs with probability $p(1-q) + q(1-r)$. In the former, it is easy to verify her belief remains unaffected, i.e., $\pi_{i,1} = \pi_{i,0}$. In the latter case, applying Bayes' rule we see:

$$\pi_{i,1} = \frac{(1-p)^2 \pi_{i,0}}{(1-p)^2 \pi_{i,0} + (1-(1-p)^2)(1-\pi_{i,0})}$$

Thus, the expected belief update of agent i is given by:

$$\mathbb{E}[\pi_{i,1}|\pi_{i,0}] = ((1-p)(1-q)+qr)\pi_{i,0} + (p(1-q)+q(1-r))\frac{(1-p)^2 \pi_{i,0}}{(1-p)^2 \pi_{i,0} + (1-(1-p)^2)(1-\pi_{i,0})}$$

Note that the change in belief from targeting, given by $\Delta \equiv \mathbb{E}[\pi_{i,1}|\pi_{i,0}] - \pi_{i,0}$ is:

$$\frac{\partial \Delta}{\partial \pi_{i,0}} = p\left(\frac{(2-p)(1-p)^2(p(1-q)+q(1-r))}{(\pi_{i,0}(2p^2-4p+1)+(2-p)p)^2} - (1-q)\right) - q(1-r)$$

When $p > 1/2$, note that $2p^2 - 4p + 1 < 0$, so $\partial\Delta/\partial\pi_{i,0}$ is strictly increasing in $\pi_{i,0}$, and thus Δ is strictly convex is $\pi_{i,0}$. Moreover, $\Delta(\pi_{i,0} = 0) = 0$ and $\Delta(\pi_{i,0} = 1) = 0$, so Δ is maximized at some unique $\pi^*_{i,0} \in (0,1)$. Note that when $\pi_{i,0} = 1/2$, then

$$\frac{\partial \Delta}{\partial \pi_{i,0}}(\pi_{i,0} = 1/2) = -(1-2(2-p)p)^2(p(1-q)+q(1-r)) < 0$$

Because $\Delta'' > 0$ everywhere, this implies that $\Delta'(\pi^*_{i,0}(r)) = 0$ for some $\pi^*_{i,0}(r) > 1/2$, which might depend on r.

Consider the set of agents with belief $\alpha = \pi_{i,0}$, denoted by $\mathcal{A}$, who are targeted. Recall that all DeGroots converge to a consensus belief π_∞:

$$\pi_\infty(r) = \int_{\alpha\in\mathcal{A}} \mathbb{E}[\pi_{i,1} \,|\, \alpha\,,\, m^p = L]\, h(\alpha)\, d\alpha + \int_{\alpha\notin\mathcal{A}} \mathbb{E}[\pi_{i,1} \,|\, \alpha\,,\, m^p = \varnothing]\, h(\alpha)\, d\alpha$$

If $\mathcal{A}$ is restricted to have some small measure ν (in prior space h), and the objective is to minimize π_∞ it is clear that the optimal choice of $\mathcal{A}$ is top pick an open interval around $\pi^*_{i,0}(r)$ given that Δ is continuous in $\pi_{i,0}$.

Finally, note that $\pi_\infty(r)$, under the optimal choice of $\mathcal{A}$ for each r, is continuous due to

Berge's theorem of the maximum. Finally, because of continuity and the fact $\pi_\infty(r) > 1/2$ for all r, we know there exists an interval $(\underline{\pi}_0^*, \bar{\pi}_0^*)$ such that for all r, π_0^* lies in this interval, with $1/2 < \underline{\pi}_0^* < \bar{\pi}_0^* < 1$. The probability that DeGroot agents mislearn is given by $\mathbb{E}_r[\mathbf{1}_{\pi_\infty(r)>1/2}]$, and because $\mathbf{1}_{\pi_\infty(r)>1/2}$ is monotone in $\pi_\infty(r)$, this implies that the optimal choice of π_0^* to maximize this expectation also satisfies $\pi_0^* > 1/2$.

(ii) *Bayesian society*: We claim that reducing the highest belief $\pi_{i,0}$ is equivalent to decreasing the likelihood of mislearning. Observe that $\pi_{i,2}$ (which is equal to $\pi_{i,T}$ for large T) is equivalent to decreasing the likelihood of mislearning for Bayesian agents. To see this, note that for Bayesian agent i, $\pi_{i,2}$ is strictly increasing in $\pi_{i,0}$, fixing the messages $\{m_i\}_{i=1}^N$, which all Bayesians are able to deduce by period 2. Note that if there exists an open interval $(\pi^1, \pi^2) \subset [\underline{\pi}, \bar{\pi}]$ such that all agents with $\pi_{i,0} \in (\pi^1, \pi^2)$ mislearn, the Bayesian society mislearns. By the assumption that $[\underline{\pi}, \bar{\pi}]$ has full support, targeting some open interval nearest $\underline{\pi}$ maximizes the probability of all agents learning when $\theta = L$. Thus, the optimal policy targets an agent who is most extreme near $\underline{\pi}$. ∎

B.3.2 Supplemental Material

Likelihood of Mislearning

We show that Bayesian agents perform worse than DeGroot agents on average. We adopt the environment from Setting B in Section 3, but we do not focus on a specific realization of the misinformation split. Instead, we look at the *likelihood* that society does not learn when we draw the split r from its true (uniform, in this example) distribution. How do Bayesian agents perform relative to DeGroot agents *on average*?

We track the beliefs of both societies:

Bayesian Population. At $t = 1$, each Bayesian agent i forms a posterior belief $\pi_{i,1}$ based on m_i according to Section 2. Once again, we know that $\pi_{j,1} > \pi_{j,0}$ if and only if $m_i = R$ and $\pi_{j,1} < \pi_{j,0}$ if and only if $m_i = L$. So every agent i can deduce all of the messages $\{m_j\}_{j=1}^N$ by period 2.

Among the collection of messages $\{m_j\}_{j=1}^N$, if the state is $\theta = L$, there are (roughly) $p(1-q) + q(1-r)$ proportion of L messages and $(1-p)(1-q) + qr$ proportion of R messages, whereas if the state is $\theta = R$, there are $(1-p)(1-q) + q(1-r)$ proportion of L messages and $p(1-q) + qr$ proportion of R messages. The state cannot be pinned down if there exists a value $r' \in [0, 1]$

such that $p(1-q)+q(1-r)=(1-p)(1-q)+q(1-r')$; in this case, there exist *exactly* two realizations of r (the true r and another r') for which the given distribution of messages can be explained under two different states, $\theta = L$ (correct) and $\theta = R$ (incorrect). Moreover, because r is uniformly distributed on $[0,1]$, *both* of these scenarios are equally likely. This implies that $\pi_{j,2} = \pi_{j,0}$ because the messages provide no information about the state θ; consequently, the society of Bayesian agents does not learn.

Note that $(1-p)(1-q)+q(1-r')$ is maximized when $r'=0$; thus it is sufficient to consider for what values of r the inequality $p(1-q)+q(1-r) \leq (1-p)(1-q)+q$ holds; this corresponds to the values of r for which there is mislearning. Rearranging, we see that $r \geq 1 - \frac{1-2p(1-q)}{q} = 0.6$ when $p = 0.6$ and $q = 0.25$. Hence, the Bayesian society mislearns with probability 40% (given that r is uniformly distributed on $[0,1]$) in the setting with 25% misinformation.

DeGroot Population. DeGroot agents update in the same way as before (i.e., via Equations (1) and (2)) using $p = 0.6$. Thus, noting that $p(1-q)+q(1-r)$ of the messages are L and $(1-p)(1-q)+qr$ of the messages are R, DeGroot agents hold beliefs $\pi_{j,t}$ for all $t \geq 2$:

$$\pi_\infty(r) \equiv \int_0^1 \left((.4(.75)+.25r) \cdot \frac{.6\alpha}{.6\alpha + .4(1-\alpha)} + (.6(.75)+.25(1-r)) \cdot \frac{.4\alpha}{.4\alpha+.6(1-\alpha)} \right) d\alpha$$

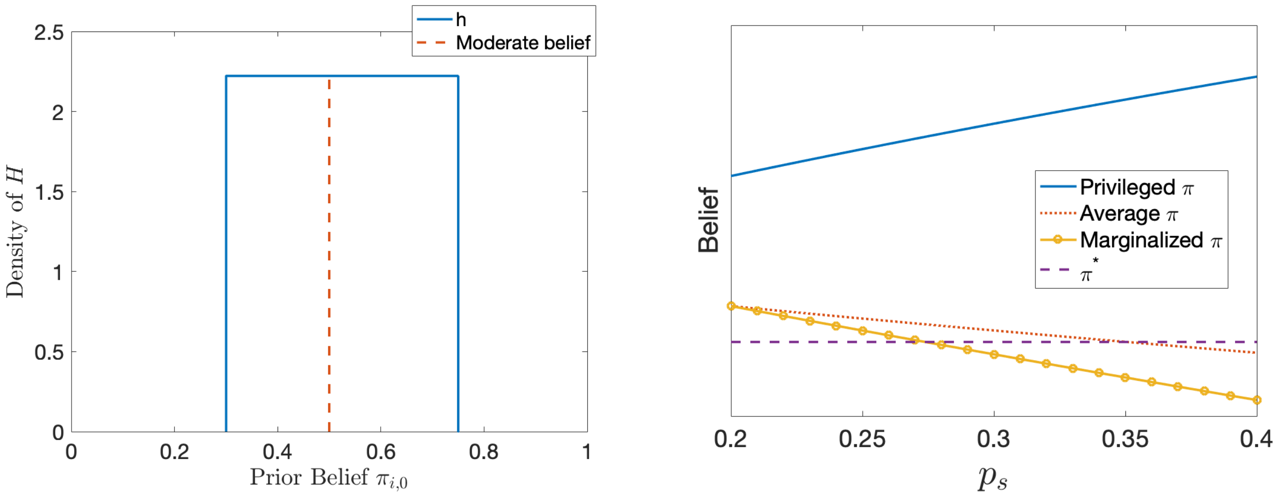
Note that $\pi_\infty(r)$ is monotonically increasing in r and it can be shown that $\pi_\infty(r) \leq 1/2$ if and only if there are less than 50% R messages; thus, $\pi_\infty(r) \leq 1/2$ if and only if $(1-p)(1-q)+qr \leq 1/2$, or in other words, $r \leq \frac{1/2-(1-p)(1-q)}{q} = 0.8$. Thus, since DeGroot agents mislearn the true state only when $r \geq r^*$, we see they mislearn 20% of the time, which *outperforms* the Bayesian population by a factor of two. Recall that quantifying how much better DeGroots do than Bayesians in general is formally analyzed in Theorem 2 in the paper.

Polarization

The next example shows that in some cases, mean-preserving spreads (i.e., more polarization) can *improve* the learning outcomes of DeGroot agents.

Example B.3.1. Consider a world where $\theta = L$ and as in Section 3, the misinformation is $q = 0.25$, so DeGroot societies (as well as Bayesian societies) mislearn with positive probability.

First, suppose H is distributed with a small right bias, as demonstrated in Figure B-17a and Figure B-17b. In the more polarized society of Figure B-17a, many prior opinions start off initially quite misinformed, so not much misinformation on the right-side can support learning (i.e., the realization of r must be lower); in particular, $r \leq .281 \equiv r^*$ is required to support learning. Next, consider a decrease in polarization to the Dirac-delta function on the average opinion of H, as shown in Figure B-17b. This increases the threshold of right-leaning misinformation that can be tolerated for learning to $r^* = .319$, and the corresponding probability that learning occurs also increases. Because moderate right-leaning agents are the most likely to be influenced by organic left-leaning news, less polarization helps the DeGroot society learn, as is the case with the Bayesians.

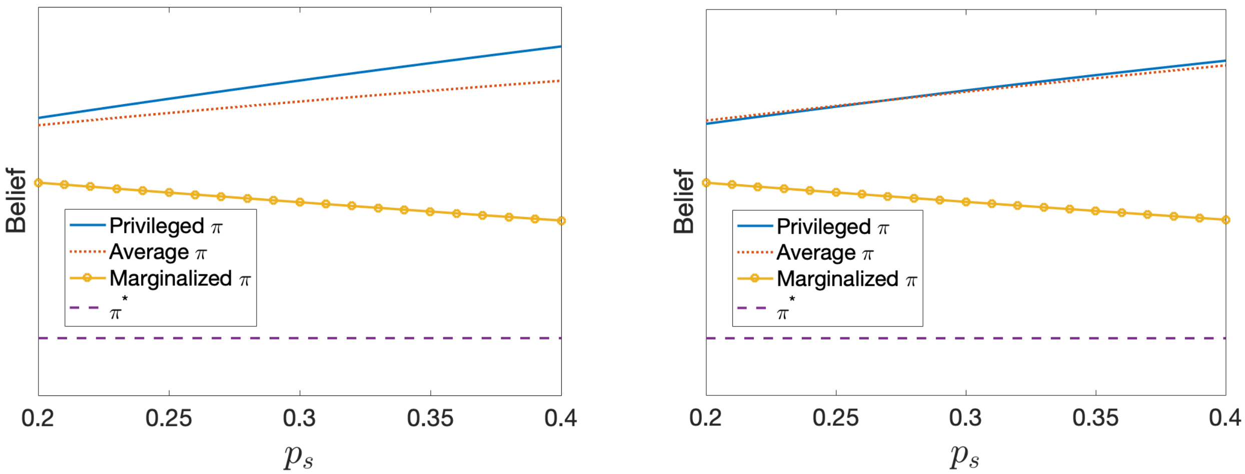


(a) Right-leaning but polarized society.

(b) Right-leaning but non-polarized society.

Figure B-17. Two right-leaning distributions of prior beliefs (with the same mean belief), one of which is polarized and the other is not. The less polarized community mislearns less often because less evidence is needed to convince moderate right-leaning agents of $\theta = L$.

Conversely, suppose H is distributed with a small left bias, as demonstrated in Figure B-18a and Figure B-18b, so beliefs tend to support the correct state of the world. When (almost all) beliefs are concentrated just left of center (Figure B-18b), there is positive probability of mislearning because there is some chance that right-leaning misinformation dominates and causes all agents to move closer to right-leaning ideas. However, these effects are mitigated when polarization increases, and learning occurs with probability 1 when it is sufficiently high, as shown in Figure B-18a. This is because strong left-leaning believers are not swayed as much by right-leaning misinformation, but moderate right-leaning agents *can* be considerably convinced of left-leaning organic news. Thus, polarization generally helps when then initial belief distribution is already slanted toward the correct state. This is somewhat surprising



(a) Left-leaning but polarized society.

(b) Left-leaning but non-polarized society.

Figure B-18. Two left-leaning distributions of prior beliefs (with the same mean belief), one of which is polarized and the other is not. The less polarized community mislearns *more often* because they are more susceptible to believing right-leaning ideas when right-leaning misinformation is more pervasive.

given that additional polarization pushes more agents toward the incorrect belief of the world. Observe that this stands in contrast to polarization in Bayesian societies: whereas DeGroots always reach a consensus, and thus polarization can sometimes nudge the entire society closer to truth (in the aggregate), the ability of Bayesians to spin multiple narrative impedes consensus in the face of increasing polarization. ▮

Mislearning Rates with High Misinformation

Let $q > q^*$. We denote by μ the ratio of the probability of DeGroot mislearning to the probability of Bayesian mislearning. Recall $\mu < 1$ by Theorem 1 (i.e., the "relative" frequency of DeGroot to Bayesian mislearning). Low values of μ indicate DeGroots do much better than Bayesians, whereas relatively larger values indicate Bayesians close the gap in mislearning more. Next, we try to reason about the conditions under which μ is increasing (i.e., Bayesian agents start picking up an advantage relative to DeGroot agents) or μ is decreasing (i.e., DeGroots learn more frequently relative to Bayesians) as a function of the strength of organic signals p and the misinformation in the system q. For, this we make the following standard definition of the *hazard rate* of a distribution:

Definition B.3.1. The *hazard rate* $\lambda_G(\alpha)$ of a distribution G is given by $\lambda_G(\alpha) = \frac{g(\alpha)}{1-G(\alpha)}$ where

$g(\cdot)$ is the density of cumulative distribution function G.

The hazard rate at point α^* corresponds the *likelihood* of the realization $\alpha \in (\alpha^*, \alpha^* + d\alpha)$ relative to the interval size $d\alpha$, conditional on $\alpha \geq \alpha^*$. Theorem 2 relates the hazard rate at specific points on the F distribution (local properties of F) to the sensitivity of μ to p (global property of F). To gather some intuition for Theorem 2, let us look at three applications:

(i) *Uniform distribution*: Recall the uniform distribution we assumed for F in Section 3 showed that when $q = 0.25$, DeGroot agents mislearn half as often as the Bayesians did. How does this depend on p? The hazard rate is $\lambda_F(r) = \frac{1}{1-r}$, which is increasing in r because the likelihood of falling within an interval of fixed length dr is increasing as one conditions on higher values of r. Thus, while $\lambda_F(\alpha - \beta) < \lambda_F(\alpha)$, it is unclear its relation to $\lambda_F(\alpha)/2$. Some basic algebra reveals that $\lambda_F(\alpha - \beta) = \lambda_F(\alpha)/2$ for *all* values of α, β, so μ has no dependence on p. Thus, DeGroot societies always mislearn half as often as Bayesian ones on the uniform distribution.

(ii) *Unskewed misinformation (Figure 2 in the paper)*: This distribution of F is one where misinformation is likely to evenly balanced between L and R. Instead of computing the hazard rate for α and $\alpha - \beta$ explicitly, we will draw inferences by comparing it to the uniform distribution. When $r < 1/2$, the hazard rate is given by $\lambda_F(r) = \frac{4r}{1-2r^2}$ whereas when $r > 1/2$, the hazard rate is given by $\lambda_F(r) = \frac{4-4r}{2-4r+2r^2}$. It is easy to show the ratio of the hazard rate of this distribution to the uniform distribution is increasing on $r < 1/2$ and constant on $r > 1/2$. Thus, $2\lambda_F(\alpha - \beta) \leq \lambda_F(\alpha)$ and the ratio μ is decreasing in p, meaning that DeGroots do comparatively better with more precise organic information in an inverted V-distribution (i.e., the Bayesians are more than twice as likely to mislearn than the Bayesians). The intuition is simple: more moderate misinformation increases the likelihood that Bayesian agents can spin a narrative to their liking, whereas for DeGroots it corresponds to a greater likelihood of having balanced misinformation (that washes out), allowing the organic news to win out.

(iii) *Skewed misinformation (inverse of Figure 2 in the paper)*: Relative to application (ii), the opposite effect occurs here. When $r < 1/2$, the hazard rate is decreasing relative to the uniform distribution; when $r > 1/2$, the hazard rate is again constant. This means the opposite inequality holds (i.e., $\lambda_F(\alpha) \leq 2\lambda_F(\alpha - \beta)$) and by Theorem 2, the ratio μ is increasing in p. When the misinformation is more extreme, Bayesians are comparatively more resilient, and mislearn less than twice as often as their DeGroot counterparts. The high likelihood of very misleading misinformation is not as well-handled by the DeGroot agents relative to a Bayesian society. While the Bayesian society can use more extreme misinformation to dismiss an incorrect narrative, the DeGroot society falls victim to such misinformation.

Network Learning Dynamics and Multiple Messages

In Section 2, we considered a model of learning where all agents observe the beliefs of all other agents. However, this is often an unrealistic assumption, and there is a wide array of literature that considers the subtleties of learning when these observations are incomplete (see Golub and Sadler (2017) for a survey). The common approach to modeling this incompleteness is to assume there is a social network with pairwise connections that determines who can observe (or talk to) whom. In this context, the model in Section 2 assumes a *complete* social network, which simplifies the relevant dynamics to two periods.

In this section, we relax this assumption by considering arbitrary network architectures and the richer learning dynamics that occur over a longer time horizon. Under relatively mild conditions on the network structure, we show network learning leads to the same outcomes and insights found in the more parsimonious complete network setting, thereby rendering our assumption to be largely without loss of generality. We do this by building off of the previous literature on network learning in both Bayesian and DeGroot populations.

Network Preliminaries. We assume that all agents are arranged in an undirected social network $\mathbf{G}$. A link $i \leftrightarrow j$ denotes that agent i and agent j observe (or talk to) each other. We let $\mathcal{N}_i$ denote the neighborhood of agent i (i.e., the set of agents j with $i \leftrightarrow j$). The adjacency matrix $\mathbf{A}$ of $\mathbf{G}$ is a binary matrix with $[\mathbf{A}]_{ii} = 1$ and $[\mathbf{A}]_{ij} = 1$ if and only if $i \leftrightarrow j$. Let $d_i^{\mathbf{G}}$ be the degree of agent i in $\mathbf{G}$. We say a network $\mathbf{G}$ is k-regular if all agents have degree k.

We consider a discrete time model (as before) but with a much longer learning horizon T, $t = 0, 1, 2, \ldots, T$. We let $\tilde{\pi}_{i,t}$ denote the belief of agent i at time t under network learning, whereas $\pi_{i,0}, \pi_{i,1}$, and $\pi_{i,2}$ denote the beliefs of agent i at time $0, 1$, and 2, respectively, in the baseline model (i.e., a complete network).

Bayesian Population. Network learning in settings with fully rational (i.e., Bayesian) agents has been studied in many contexts, most notably in Acemoglu et al. (2011) and Gale and Kariv (2003). As is common in many models of Bayesian network learning,[12] we assume that the network $\mathbf{G}$ and initial priors $\pi_{i,0}$ are common knowledge.[13] Bayesian agents observe the beliefs of all agents in their neighborhoods for all $t \geq 1$ (i.e., agent i observes at time t the beliefs from $t-1$, $\{\pi_{j,t-1}\}_{j \in \mathcal{N}_i}$). Our next result shows that terminal beliefs in network learning indeed converge to the terminal beliefs of the baseline model:

Claim B.3.1. *Suppose* $\mathbf{G}$ *is connected. Then as* $T \to \infty$, $\tilde{\pi}_{i,T} \to \pi_{i,2}$.

This claim follows directly from Mueller-Frank (2013). While agents do not hold a common prior about θ, common knowledge of the heterogenous priors $\{\pi_{j,0}\}_{j=1}^{N}$ allows agents to recalibrate the (updated) beliefs they see to their own prior. It is clear that the private information at $t = 1$ (i.e., the messages) are drawn from a finite partition of the θ state space (conditional on misinformation split r). Thus, by Theorem 4 of Mueller-Frank (2013), all Bayesian agents uncover the private information (i.e., $t = 1$ messages) of all other agents (including non-neighbors) in the network as $T \to \infty$, as is the case at $t = 2$ in the baseline model.

DeGroot Population. Due to demanding assumptions about the reasoning abilities of Bayesian agents, "rule-of-thumb" learning has become a popular alternative model. The most common model is that of Degroot (1974), and later expanded upon in works such as Golub and Jackson (2010) and DeMarzo et al. (2003). In these models, agents are assumed to update their beliefs using the simple heuristic of taking linear combinations of their neighbors' beliefs. Formally, agent i forms belief $\pi_{i,t+1}$ at each time t by computing:

$$\pi_{i,t+1} = \frac{1}{1 + d_i^{\mathbf{G}}} \left(\pi_{i,t} + \sum_{j \in \mathcal{N}_i} \pi_{j,t} \right)$$

Our next result provides conditions under which DeGroot learning over the network $\mathbf{G}$ leads to the same terminal beliefs as in our baseline model:

[12] In addition to Acemoglu et al. (2011) and Gale and Kariv (2003), see Mueller-Frank (2014) and Mossel et al. (2014).

[13] An alternative assumption, which does not require strong common knowledge assumptions of non-neighbor priors or the network structure, is that the size of the smallest neighborhood grows unboundedly as $N \to \infty$.

Claim B.3.2. *Suppose* $\mathbf{G}$ *is a connected,* k*-regular network. Then as* $T \to \infty$*,* $\tilde{\pi}_{i,T} \to \pi_{i,2}$*.*

This claim follows directly from Golub and Jackson (2010). First, by Proposition 1 in Golub and Jackson (2010), observe that consensus is reached (as in the baseline model) because the normalized adjacency matrix $\mathbf{A}$ is irreducible and aperiodic, the former following from the connectedness assumption and the latter following from a positive diagonal on $\mathbf{A}$. Second, by Theorem 3 in Golub and Jackson (2010), the consensus belief of the agents as $T \to \infty$ is given by $\tilde{\pi}_{i,\infty} = \sum_{j=1}^{N} v_j^{\mathbf{G}} \pi_{j,1}$ for all agents i, where $v_j^{\mathbf{G}}$ is the (eigenvector) centrality of agent j (according to the row-stochastic normalized adjacency matrix $\mathbf{A}$). Because $v_j^{\mathbf{G}} = d_j^{\mathbf{G}} / \sum_{\ell=1}^{N} d_\ell^{\mathbf{G}}$, we obtain by k-regularity that $\tilde{\pi}_{i,\infty} = \frac{1}{N} \sum_{j=1}^{N} \pi_{j,1} = \pi_{i,2}$.

Observe that Claim B.3.2 requires an additional condition not present in Claim B.3.1, which is that no agent is more "influential" than any other agent in the network $\mathbf{G}$, as measured by her degree. This is easily satisfied by many network topologies, including several classes of random networks such as Erdos-Renyi networks (where links between agents occur uniformly at random).[14]

Multiple Messages. Let us consider the complete network setting of Section 2 for simplicity, but note that the reduction from arbitrary network learning discussed previously still applies.

In a Bayesian society with $N \to \infty$, by the strong law of large numbers, the first round of messages reveals the true fraction of R messages, ρ_R, and the true fraction of L messages, ρ_L, almost surely. Obtaining additional messages in subsequent rounds does not alter the (almost surely) known values of ρ_R or ρ_L, thus, learning is entirely unaffected by more incoming messages.

In a DeGroot society, after the first round of messages, agents converge to a consensus about θ which is a function of ρ_R (and ρ_L) alone. When H is symmetric, whether $\rho_R > 1/2$ or $\rho_L > 1/2$ determines if the consensus, call it π_2, lies more toward state R (i.e., $\pi_2 > 1/2$) or state L (i.e., $\pi_2 < 1/2$). By the martingale property of Bayesian updating, it is easy to see that $\mathbb{E}[\text{BU}(\pi_2) \,|\, \rho_R > 1/2\,;\, \pi_2 > 1/2] > 1/2$ and $\mathbb{E}[\text{BU}(\pi_2) \,|\, \rho_R < 1/2\,;\, \pi_2 < 1/2] < 1/2$ (where BU is the Bayesian update for DeGroot agents conditioning on the message, given by Equations (1) and (2) of the main text). Therefore, one can show by induction that beliefs remain on the same side of belief $1/2$ as they are at $t = 2$ for all $t \geq T$, even with additional messages. Consequently,

[14] k-regularity will hold approximately for large N in dense ER networks; see for instance, Avella-Medina et al. (2020) and Dasaratha (2020).

the likelihood of (mis)learning is unaffected by any further stream of messages.

Robustness: Learning Metric and Finite Populations

We conduct two robustness checks on our main learning results. First, we consider how our results change when looking at the *expected* fraction of mislearning agents, instead of the binary metric of whether the entire society learns or not. Second, we test the sensitivity of our large population assumption (i.e., $N \to \infty$) by simulating learning in settings with finite N.

Learning Metric

We consider the alternative learning metric of the expected proportion of the population that mislearns the true state. In particular, we look at (i) the environments where DeGroots or Bayesians perform better under this metric, and (ii) how the targeting policy changes under this other learning objective.

DeGroot vs Bayesian. How is learning affected when one evaluates the expected fraction of mislearning agents? Provided that H is sufficiently polarized (i.e., there are few moderate left or right-leaning agents and most agents have relatively strong opinions), the expected fraction of mislearning Bayesian agents is approximately half of the Bayesian mislearning rate; conversely, the expected fraction of mislearning DeGroot agents is exactly the mislearning rate. The former can be seen from Proposition 2: all agents in society learn when there is a single narrative, but when there are multiple narratives (and H is symmetric), only those who have priors that agree with the truth (i.e., 50%) will take the correct action. The latter can be seen from the fact that the DeGroot society always comes to consensus, so the two notions of learning coincide.

The comparison between Bayesian and DeGroot societies using this learning metric is best highlighted using the hazard rate analysis of Section 4.4 and Appendix B.3.2 in the high-misinformation regime (i.e., $q > q^*$), which depends on the distribution of misinformation F:

(i) *Uniform distribution*: When F has a uniform distribution, the Bayesian society mislearns twice as often as the DeGroot society. Thus, under the new learning metric, the expected fraction of mislearning agents is *the same* for both Bayesians and DeGroots.

(ii) *Unskewed distribution*: When F has an unskewed distribution (i.e., misinformation is more likely to be evenly split between ideologies), DeGroots perform better than Bayesians relative to the base case of the uniform distribution. Thus, under the new learning metric, DeGroots outperform Bayesian agents in the expected fraction of mislearning agents.

(iii) *Skewed distribution*: When F has a more skewed distribution (i.e., misinformation is more likely to come mostly from one ideology), Bayesians perform better than DeGroots relative to the base case of the uniform distribution. Thus, under the new learning metric, Bayesians would outperform DeGroot agents in the expected fraction of mislearning agents.

These cases make it clear that the main message of our paper —that reasoning abilities have an ambiguous effect on learning outcomes and that DeGroot agents can outperform Bayesian agents— is not an artifact of the learning definition but rather a fundamental property of learning in the presence of misinformation. Under this metric, the outcomes depend on both the level of misinformation (i.e., whether $q > q^*$ or $q < q^*$) and the concavity/convexity of f. An interesting implication of the above metric is that it suggests that DeGroot agents are better at learning the state when misinformation is likely to come from both sides of the spectrum, which is likely the case with most controversial political issues. On the other hand, Bayesian agents can be better at learning the state when misinformation is mostly one-sided, e.g., that it argues for the earth being flat.

Targeting Policies. We consider a targeting policy where the regulator wants to minimize the expected proportion of agents who mislearn. As before, we assume $q > q^*$ and the true state is $\theta = L$, which is known to the regulator. Because DeGroot agents always converge to a consensus, the regulator does not change her targeting policy because either all agents learn or none do; we let π^*_D denote the belief of the optimal DeGroot target which is the same as π^* in Proposition 3. However, the policy will change for the Bayesian society in a subtle way. For the most likely split of misinformation that generates two narratives, the regulator targets the agent whose posterior belief is barely to the right of belief $1/2$. We illustrate how this targeting policy works in practice for the three settings considered before:

(i) *Uniform distribution*: For the uniform distribution, when there are two narratives, no agents change their prior beliefs (and when there is only one, all agents learn anyway).

Therefore, the optimal targeting policy is to target the most moderate right-leaning agent. This stands in contraposition to Proposition 3, as the regulator is better off targeting more moderate Bayesian agents and relatively more extreme DeGroots (i.e., $p_B^* < p_D^*$).

(ii) *Unskewed distribution*: When F follows the inverse-V distribution, there are two cases to consider. The first case is that the misinformation split $r = 1/2$ admits two narratives; this occurs when $q > \frac{2(2p-1)}{4p-1} > q^*$. In this case, there is some $r^* < 1/2$ that explains the $\theta = R$ narrative; however, this narrative is less compelling than the $\theta = L$ narrative (because f is largest at $r = 1/2$). Thus, some moderate right-leaning agent with belief $\pi_B^* > 1/2$ will be the optimal target. As q increases, the r^* corresponding to the R narrative moves closer to $1/2$ and thus becomes more likely. Thus, for low q, π_B^* will be close to 1 (target the extremists, as in Proposition 3) whereas for high q, π_B^* will target the most right-leaning moderates (as in the uniform distribution).

The second case is that the misinformation split $r = 1/2$ admits only one narrative ($q^* < q < \frac{2(2p-1)}{4p-1}$). Then there are (barely) two narratives when $r = r^*$ satisfies $(1-p)(1-q) + qr^* = p(1-q)$, or $r^* = \frac{(2p-1)(1-q)}{q} > 1/2$, and this is the most likely split of misinformation conditional on the existence of two narratives. The other narrative (the $\theta = R$ narrative) has likelihood almost zero (this narrative occurs at $r = 0$), so only the most polarized right-leaning agents will prescribe to this narrative over the true narrative for $\theta = L$. Thus, the optimal policy is exactly the same as in Proposition 3: the regulator should target the most extreme Bayesian agents, which are strictly more extreme than the optimal DeGroot target.

(iii) *Skewed distribution*: When F follows the V-distribution, the most likely split of misinformation that admits two narratives is $r = 1$ (note that $r = 0$ is also the most likely, but when $\theta = L$, the L narrative is unique). There are always two narratives in this case and the $\theta = R$ narrative corresponds to $r^* = \frac{1-2p(1-q)}{q}$. When $q > \frac{2(2p-1)}{4p-1} > q^*$, $r^* > 1/2$, so increasing q increases r^* and makes the $\theta = R$ narrative more likely. Consequently, π_B^* decreases and the regulator targets more moderate right-leaning agents. When $q^* < q < \frac{2(2p-1)}{4p-1}$, $r^* < 1/2$, so increasing q increases r^* but makes the $\theta = R$ narrative *less* likely. Consequently, π_B^* *increases* and the regulator targets more extreme right-leaning agents. In particular, we recover the policy of Proposition 3 (target the extremists) when q has the intermediate value of $\frac{2(2p-1)}{4p-1}$.

Discussion — These cases highlight an interesting feature of the Bayesian targeting policy under this alternative learning metric: the optimal target depends on the quantity of misinformation q and can even be *non-monotone* in q.

In case (ii) (as with any concave distribution for F), with relatively low misinformation, the policy is the same as in Proposition 3: the regulator should target extremists because these are the most stubborn agents to convince. However, with relatively high misinformation, the policy changes to focus on more moderate right-leaning agents, because abundant misinformation will necessarily confound learning for the extremists.

In case (iii) (as with any convex distribution for F), the regulator employs a non-monotone targeting policy. When misinformation is small or large, both narratives conclude that misinformation is heavily skewed toward one ideology, which inhibits learning and admits an optimal policy of targeting moderate right-leaning agents. However, when misinformation is moderate, the incorrect narrative involves an even split of misinformation, which is unlikely; consequently, the regulator should target the most difficult agents to convince, the extremists, as in Proposition 3.

Finite Populations

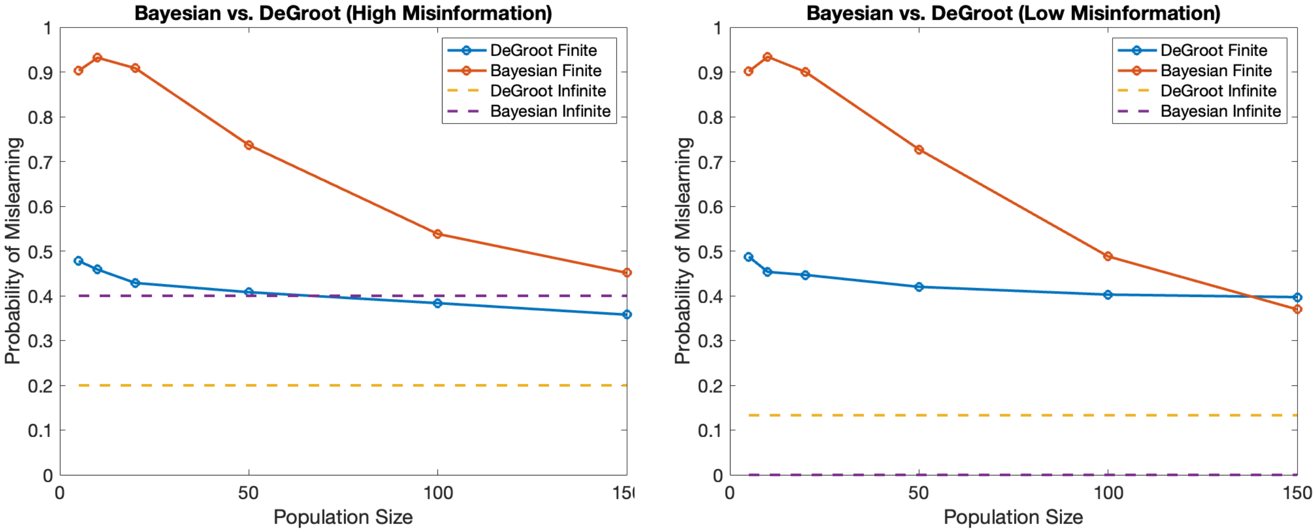
Recall that Theorem 1 shows that when $N \to \infty$ and in settings where misinformation is high ($q > q^*$), DeGroot societies outperform Bayesian societies, whereas the converse holds when misinformation is low ($q < q^*$). We consider how robust these findings are to a *finite* population of agents.

High Misinformation. When there is high misinformation, DeGroot agents consistently outperform Bayesian agents for all finite populations, as shown in Figure B-19a. Notably, Bayesian agents converge to their theoretical long-run mislearning average more quickly than DeGroot agents.[15] With few agents in the population, the small amount of information on the state θ allows for more wild narrative telling. We denote this kind of narrative telling as a *noise-based narrative* to distinguish it from the more subtle narrative telling identified in Section 4.2. Noise-based narratives arise from the agents believing that the existing (and

[15] While both Figure B-19a and Figure B-19b show DeGroots bounded away from their long-run average, one can verify via simulation that after $N > 10000$, DeGroots are within 1% of their $N \to \infty$ learning rate (i.e., convergence is slower). The plots are capped at $N \leq 150$ because of the numerical issues associated with Bayesian learning when N is large but finite.

limited) content is just by happenstance in opposition to one's priors, but is not indicative of θ. Once the population reaches a critical mass, the noise-based narrative dwindles and only the misinformation narrative spin identified in Section 4.2 persists, keeping Bayesian mislearning at or above 40%.

Low Misinformation. When there is low misinformation, DeGroot agents can still outperform Bayesian agents in small populations, as shown in Figure B-19b. This is again related to the noise-based narrative that Bayesian agents can spin in finite populations. As seen in Figure B-19b, with an infinite population, the Bayesian society learns almost surely, and so there are no misinformation narratives (of the form in Section 4.2) that can be told. As the population increases, the noise-based narratives attenuate and once the population is large enough (e.g., $N > 150$), the Bayesians begin to outperform the DeGroots, as predicted in Theorem 1 for the low misinformation regime.



(a) High misinformation regime

(b) Low misinformation regime

Figure B-19. Beliefs in finite populations

Bibliography

Abramowitz, Alan I. (2010), *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*. Yale University Press.

Abreu, Jorge, João Nogueira, Valdecir Becker, and Bernardo Cardoso (2017), "Survey of Catch-up TV and other time-shift services: a comprehensive analysis and taxonomy of linear and nonlinear television." *Telecommunication Systems*, 64, 57–74.

Acemoglu, Daron, Victor Chernozhukov, and Muhamet Yildiz (2016), "Fragility of asymptotic agreement under bayesian learning." *Theoretical Economics*, 11, 187–225.

Acemoglu, Daron, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar (2013), "Opinion fluctuations and disagreement in social networks." *Mathematics of Operations Research*, 38, 1–27.

Acemoglu, Daron, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar (2011), "Bayesian learning in social networks." *The Review of Economic Studies*, 78, 1201–1236.

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar (2022a), "Too Much Data: Prices and Inefficiencies in Data Markets." *American Economic Journal: Microeconomics (forthcoming)*.

Acemoglu, Daron, Asuman Ozdaglar, and Ali ParandehGheibi (2010), "Spread of (mis)information in social networks." *Games and Economic Behavior*, 70, 194–227.

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius (2022b), "A model of online misinformation." Technical report, National Bureau of Economic Research.

Acemoglu, Daron, Asuman Ozdaglar, and Alireza Tahbaz-Salehi (2015), "Systemic Risk and Stability in Financial Networks." *American Economic Review*, 105, 564–608.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva (2020), "The polarization of reality." In *AEA Papers and Proceedings*, volume 110, 324–28.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow (2020), "The Welfare Effects of Social Media." *American Economic Review*, 110, 629–676.

Allcott, Hunt and Matthew Gentzkow (2017), "Social media and fake news in the 2016 election." *Journal of Economic Perspectives*, 31, 211–36.

Allcott, Hunt, Matthew Gentzkow, and Lena Song (2022), "Digital Addiction." *American Economic Review*, 112, 2424–2463.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts (2020), "Evaluating the fake news problem at the scale of the information ecosystem." *Science Advances*, 6.

Allon, Gad, Kimon Drakopoulos, and Vahideh Manshadi (2021), "Information Inundation on Platforms and Implications." *Operations Research*, 69, 1784–1804.

Allon, Gad and Dennis Zhang (2017), "Managing service systems in the presence of social networks." *Available at SSRN 2673137*.

Altay, Sacha, Anne-Sophie Hacquin, and Hugo Mercier (2020), "Why do so few people share fake news? It hurts their reputation." *New Media & Society*.

Andrews, Lori (2012), "Facebook is using you." *The New York Times*, 4.

Apprich, Clemens, Florian Cramer, Wendy Hui Kyong Chun, and Hito Steyerl (2018), *Pattern Discrimination*. University of Minnesota Press.

Aral, Sinan and Paramveer S. Dhillon (2018), "Social influence maximization under empirical influence models." *Nature Human Behaviour*, 2, 375–382.

Arditi, David (2021), *Streaming Culture: Subscription Platforms And The Unending Consumption Of Culture*. Emerald Group Publishing. Google-Books-ID: MXAnEAAAQBAJ.

Arnold, Barry C. (1987), *Majorization and the Lorenz Order: A Brief Introduction*. Lecture Notes in Statistics, Springer-Verlag, New York.

Avella-Medina, Marco, Francesca Parise, Michael T. Schaub, and Santiago Segarra (2020), "Centrality measures for graphons: Accounting for uncertainty in networks." *IEEE Transactions on Network Science and Engineering*, 7, 520–537.

Babus, Ana (2016), "The formation of financial networks." *The RAND Journal of Economics*, 47, 239–272.

Bago, Bence, David G Rand, and Gordon Pennycook (2020), "Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines." *Journal of experimental psychology: general*.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic (2015), "Exposure to ideologically diverse news and opinion on Facebook." *Science*, 348, 1130–1132.

Banaji, Shakuntala, Ramnath Bhat, Anushi Agarwal, Nihal Passanha, and Mukti Sadhana Pravin (2019), "WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India." *Department of Media and Communications, London School of Economics*.

Banerjee, Abhijit V. (1992), "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics*, 107, 797–817.

Bashir, Hilal and Shabir Ahmad Bhat (2017), "Effects of social media on mental health: A review." *International Journal of Indian Psychology*, 4, 125–131.

Bennett, Colin and Jesse Gordon (2020), "Understanding the "Micro" in Micro-Targeting: An Analysis of Facebook Digital Advertising in the 2019 Federal Canadian Election." *Available at SSRN 3589687*.

Berger, Benedikt, Christian Matt, Dennis M. Steininger, and Thomas Hess (2015), "It Is Not Just About Competition with "Free": Differences Between Content Formats in Consumer Preferences and Willingness to Pay." *Journal of Management Information Systems*, 32, 105–128.

Berryman, Chloe, Christopher J. Ferguson, and Charles Negy (2018), "Social media use and mental health among young adults." *Psychiatric quarterly*, 89, 307–314.

Bickert, Monika (2020), "Charting a Way Forward on Online Content Regulation."

Bikhchandani, Sushil, David Hirshleifer, Omer Tamuz, and Ivo Welch (2021), "Information Cascades and Social Learning." Working Paper 28887, National Bureau of Economic Research.

Bohren, J Aislinn and Daniel N Hauser (2017), "Bounded rationality and learning: A framework and a robustness result." *PIER Working Paper*.

Breza, Emily, Arun G. Chandrasekhar, and Alireza Tahbaz-Salehi (2018), "Seeing the forest for the trees? An investigation of network knowledge." *arXiv:1802.08194 [physics, stat]*.

Broniatowski, David A, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze (2018), "Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate." *American journal of public health*, 108, 1378–1384.

Bronstein, Michael V, Gordon Pennycook, Adam Bear, David G Rand, and Tyrone D Cannon (2019), "Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking." *Journal of applied research in memory and cognition*, 8, 108–117.

Buchanan, Tom (2020), "Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation." *PLOS ONE*, 15.

Budak, Ceren, Divyakant Agrawal, and Amr El Abbadi (2011), "Limiting the spread of misinformation in social networks." In *Proceedings of the 20th international conference on World wide web*, WWW '11, 665–674.

Budzinski, Oliver, Sophia Gaenssle, and Nadine Lindstädt-Dreusicke (2021), "The battle of YouTube, TV and Netflix: an empirical analysis of competition in audiovisual media markets." *SN Business & Economics*, 1, 116.

Bursztyn, Leonardo, Aakaash Rao, Christopher P Roth, and David H Yanagizawa-Drott (2020), "Misinformation during a pandemic." Technical report, National Bureau of Economic Research.

Butters, Gerard R. (1977), "Equilibrium Distributions of Sales and Advertising Prices." *Review of Economic Studies*, 44, 465–491.

Candogan, Ozan and Kimon Drakopoulos (2020), "Optimal signaling of content accuracy: Engagement vs. misinformation." *Operations Research*, 68, 497–515.

Catlett, Natalie (2022), "Social Media Use and Mental Health: An Educational Intervention to Reduce Depression and Anxiety in Adolescents." *DNP Projects*.

Cen, Sarah H and Devavrat Shah (2020), "Regulating algorithmic filtering on social media." *arXiv preprint arXiv:2006.09647*.

Centola, Damon (2010), "The spread of behavior in an online social network experiment." *science*, 329, 1194–1197.

Centola, Damon and Michael Macy (2007), "Complex Contagions and the Weakness of Long Ties." *American Journal of Sociology*, 113, 702–734.

Chen, Li and Yiangos Papanastasiou (2021), "Seeding the Herd: Pricing and Welfare Effects of Social Learning Manipulation." *Management Science*, 67, 6734–6750.

Chyi, Hsiang Iris (2005), "Willingness to Pay for Online News: An Empirical Study on the Viability of the Subscription Model." *Journal of Media Economics*, 18, 131–142.

Chyi, Hsiang Iris and Yee Man Margaret Ng (2020), "Still Unwilling to Pay: An Empirical Analysis of 50 U.S. Newspapers' Digital Subscription Results." *Digital Journalism*, 8, 526–547.

Cinelli, Matteo, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala (2020), "The COVID-19 social media infodemic." *Scientific Reports*, 10, 16598.

Clogston, Juanita "Frankie" (2016), "The Repeal of the Fairness Doctrine and the Irony of Talk Radio: A Story of Political Entrepreneurship, Risk, and Cover." *Journal of Policy History*, 28, 375–396.

Cuthbertson, Richard, Peder Inge Furseth, and Stephen J. Ezell (2015), "Facebook and MySpace: The Importance of Social Networks." In *Innovating in a Service-Driven Economy: Insights, Application, and Practice* (Richard Cuthbertson, Peder Inge Furseth, and Stephen J. Ezell, eds.), 145–158, Palgrave Macmillan UK, London.

Dasaratha, Krishna (2020), “Distributions of centrality on networks.” *Games and Economic Behavior*, 122, 1–27.

De Jans, Steffi, Dieneke Van de Sompel, Liselot Hudders, and Veroline Cauberghe (2019), “Advertising targeting young children: an overview of 10 years of research (2006–2016).” *International Journal of Advertising*, 38, 173–206.

Degroot, Morris H. (1974), “Reaching a Consensus.” *Journal of the American Statistical Association*, 69, 118–121.

DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel (2003), “Persuasion Bias, Social Influence, and Unidimensional Opinions*.” *The Quarterly Journal of Economics*, 118, 909–968.

Deng, Yiting and Carl F. Mela (2018), “TV viewing and advertising targeting.” *Journal of Marketing Research*, 55, 99–118.

Drummond, Caitlin and Baruch Fischhoff (2017), “Individuals with greater science literacy and education have more polarized beliefs on controversial science topics.” *Proceedings of the National Academy of Sciences*, 114, 9587–9592.

Duffy, Andrew, Edson Tandoc, and Rich Ling (2020), “Too good to be true, too good not to share: the social utility of fake news.” *Information, Communication & Society*, 23, 1965–1979.

Eckles, Dean, René F. Kizilcec, and Eytan Bakshy (2016), “Estimating peer effects in networks with peer encouragement designs.” *Proceedings of the National Academy of Sciences*, 113, 7316–7322.

Edosomwan, Simeon, Sitalaskshmi Kalangot Prakasan, Doriane Kouame, Jonelle Watson, and Tom Seymour (2011), “The history of social media and its impact on business.” *Journal of Applied Management and entrepreneurship*, 16, 79–91.

Egelhofer, Jana Laura and Sophie Lecheler (2019), “Fake news as a two-dimensional phenomenon: a framework and research agenda.” *Annals of the International Communication Association*, 43, 97–116.

Epstein, Ziv and Hause Lin (2022), “Yourfeed: Towards open science and interoperable systems for social media.” *arXiv preprint arXiv:2207.07478*.

Epstein, Ziv, Hause Lin, Gordon Pennycook, and David Rand (2022), “How many others have shared this? experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media.” *arXiv preprint arXiv:2207.07562*.

Esponda, Ignacio and Demian Pouzo (2016), “Berk–Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models.” *Econometrica*, 84, 1093–1130.

Fiorina, Morris P., Samuel A. Abrams, and Jeremy C. Pope (2008), “Polarization in the American Public: Misconceptions and Misreadings.” *The Journal of Politics*, 70, 556–560.

Fisher, Marc, John Woodrow Cox, and Peter Hermann (2016), “Pizzagate: From rumor, to hashtag, to gunfire in dc.” *Washington Post*.

Flew, Terry (2021), “Willingness to Pay: News Media.” *International Institute of Communication*.

Fourney, Adam, Miklos Z. Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz (2017), “Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election.” In *CIKM*, volume 17, 6–10.

Gale, Douglas and Shachar Kariv (2003), “Bayesian learning in social networks.” *Games and Economic Behavior*, 45, 329–346.

Garimella, Kiran and Dean Eckles (2020), “Images and misinformation in political groups: Evidence from WhatsApp in India.” *arXiv preprint arXiv:2005.09784*.

Gentzkow, Matthew and Jesse M. Shapiro (2006), “Media Bias and Reputation.” *Journal of Political Economy*, 114, 280–316.

Giordano, Amanda L., Lindsay A. Lundeen, Kelly L. Wester, Jaewoo Lee, Samuel Vickers, Michael K. Schmit, and In Kee Kim (2022), “Nonsuicidal Self-Injury on Instagram: Examining Hashtag Trends.” *International Journal for the Advancement of Counselling*, 44, 1–16.

Goel, Ashish and Latika Gupta (2020), “Social Media in the Times of COVID-19.” *Journal of Clinical Rheumatology*, 10.1097/RHU.0000000000001508.

Golub, Benjamin and Matthew O Jackson (2010), “Naive learning in social networks and the wisdom of crowds.” *American Economic Journal: Microeconomics*, 2, 112–49.

Golub, Benjamin and Matthew O Jackson (2012), “How homophily affects the speed of learning and best-response dynamics.” *The Quarterly Journal of Economics*, 127, 1287–1338.

Golub, Benjamin and Evan Sadler (2017), “Learning in Social Networks.” SSRN Scholarly Paper ID 2919146, Social Science Research Network, Rochester, NY.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer (2019), “Fake news on Twitter during the 2016 U.S. presidential election.” *Science*, 363, 374–378.

Grossman, Gene M. and Carl Shapiro (1984), “Informative Advertising with Differentiated Products.” *The Review of Economic Studies*, 51, 63–81.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker (2019), “Less than you think: Prevalence and predictors of fake news dissemination on Facebook.” *Science Advances*, 5.

Guess, Andrew, Brendan Nyhan, and Jason Reifler (2018), “Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign.” *European Research Council*, 9, 4.

Hamilton, Lawrence C, Joel Hartter, and Kei Saito (2015), "Trust in scientists on climate change and vaccines." *Sage Open*, 5, 2158244015602752.

Holm, Anna B. and Franziska Günzel-Jensen (2017), "Succeeding with freemium: strategies for implementation." *Journal of Business Strategy*, 38, 16–24.

Hsu, Chin-Chia, Amir Ajorlou, and Ali Jadbabaie (2020), "News Sharing, Persuasion, and Spread of Misinformation on Social Networks." *Working paper*.

Hwang, Elina H and Stephanie Lee (2021), "A nudge to credible information as a countermeasure to misinformation: Evidence from twitter." *Available at SSRN*.

Jackson, Matthew O., Tomas Rodriguez-Barraquer, and Xu Tan (2012), "Social Capital and Social Quilts: Network Patterns of Favor Exchange." *American Economic Review*, 102, 1857–1897.

Jadbabaie, Ali, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi (2012), "Non-bayesian social learning." *Games and Economic Behavior*, 76, 210–225.

Kahan, Dan M, Ellen Peters, Erica Cantrell Dawson, and Paul Slovic (2017), "Motivated numeracy and enlightened self-government." *Behavioural public policy*, 1, 54–86.

Kahan, Dan M, Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel (2012), "The polarizing impact of science literacy and numeracy on perceived climate change risks." *Nature climate change*, 2, 732–735.

Kamenica, Emir (2019), "Bayesian Persuasion and Information Design." *Annual Review of Economics*, 11, 249–272.

Kamenica, Emir and Matthew Gentzkow (2011), "Bayesian Persuasion." *American Economic Review*, 101, 2590–2615.

Kanak, Zafer (2017), "Rescuing the Financial System: Capabilities, Incentives, and Optimal Interbank Networks." Technical Report 17-17, NET Institute.

Keppo, Jussi, Michael Jong Kim, and Xinyuan Zhang (2019), "Learning Manipulation Through Information Dissemination." *Working Paper, SSRN Scholarly Paper ID 3465030*.

Kim, Dam Hee, S. Mo Jones-Jang, and Kate Kenski (2020), "Why Do People Share Political Information on Social Media?" *Digital Journalism*, 0, 1–18.

Kozyreva, Anastasia, Stephan Lewandowsky, and Ralph Hertwig (2020), "Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools." *Psychological Science in the Public Interest*, 21, 103–156.

Lazer, David MJ, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, and David Rothschild (2018), "The science of fake news." *Science*, 359, 1094–1096.

Lee, Chei Sian, Long Ma, and Dion Hoe-Lian Goh (2011), "Why Do People Share News in Social Media?" In *Active Media Technology*, Lecture Notes in Computer Science, 129–140, Springer.

Levy, Ro'ee (2021), "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment." *American Economic Review*, 111, 831–870.

Liu, Qingmin (2011), "Information acquisition and reputation dynamics." *The Review of Economic Studies*, 78, 1400–1425.

Loomba, Sahil, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson (2021), "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa." *Nature human behaviour*, 1–12.

Luca, Michael and Georgios Zervas (2016), "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." *Management Science*, 62, 3412–3427.

Marotta, Veronica, Yue Wu, Kaifu Zhang, and Alessandro Acquisti (2021), "The Welfare Impact of Targeted Advertising Technologies." *Information Systems Research*.

Marshall, Albert W., Ingram Olkin, and Barry C. Arnold (2011), *Inequalities: Theory of Majorization and Its Applications*, 2 edition. Springer Series in Statistics, Springer-Verlag, New York.

Marwick, Alice E. and Rebecca Lewis (2017), "Media manipulation and disinformation online." *Data & Society Research Institute*.

McIntyre, Lee (2018), *Post-truth*. MIT Press.

McWilliams, Jenna (2009), "How Facebook beats MySpace."

Meurer, Michael and Dale O. Stahl (1994), "Informative advertising and product match." *International Journal of Industrial Organization*, 12, 1–19.

Molina, Maria D., S. Shyam Sundar, Thai Le, and Dongwon Lee (2021), ""Fake News" Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content." *American Behavioral Scientist*, 65, 180–212.

Montag, Christian, Bernd Lachmann, Marc Herrlich, and Katharina Zweig (2019), "Addictive Features of Social Media/Messenger Platforms and Freemium Games against the Background of Psychological and Economic Theories." *International Journal of Environmental Research and Public Health*, 16, 2612.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David Rand (2021a), "Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment." *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G. Rand (2021b), “Shared partisanship dramatically increases social tie formation in a twitter field experiment.” *Proceedings of the National Academy of Sciences*, 118.

Mossel, Elchanan, Allan Sly, and Omer Tamuz (2014), “Asymptotic learning on Bayesian social networks.” *Probability Theory and Related Fields*, 158, 127–157.

Mostagir, Mohamed (2010), “Exploiting myopic learning.” In *International Workshop on Internet and Network Economics*, 306–318, Springer.

Mostagir, Mohamed, Asu Ozdaglar, and James Siderius (2022), “When is society susceptible to manipulation?” *Management Science, forthcoming*.

Mostagir, Mohamed and James Siderius (2021), “Centrality in stochastic networks.” *Working paper*.

Mostagir, Mohamed and James Siderius (2022a), “Learning in a Post-Truth World.” *Management Science*, 68, 2860–2868.

Mostagir, Mohamed and James Siderius (2022b), “Social inequality and the spread of misinformation.” *Management Science, forthcoming*.

Mostagir, Mohamed and James Siderius (2022c), “Strategic reviews.” *Management Science*.

Mueller-Frank, Manuel (2013), “A general framework for rational learning in social networks.” *Theoretical Economics*, 8, 1–40.

Mueller-Frank, Manuel (2014), “Does one bayesian make a difference?” *Journal of Economic Theory*, 154, 423–452.

Neyazi, Taberez Ahmed, Antonis Kalogeropoulos, and Rasmus K. Nielsen (2021), “Misinformation concerns and online news participation among internet users in India.” *Social Media+ Society*, 7, 20563051211009013.

Nguyen, Nam P., Guanhua Yan, My T. Thai, and Stephan Eidenbenz (2012), “Containment of misinformation spread in online social networks.” In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci ’12, 213–222.

Papakyriakopoulos, Orestis, Simon Hegelich, Morteza Shahrezaye, and Juan Carlos Medina Serrano (2018), “Social media and microtargeting: Political data processing and the consequences for Germany.” *Big Data & Society*, 5.

Papanastasiou, Yiangos (2020), “Fake News Propagation and Detection: A Sequential Model.” *Management Science*, 66, 1826–1846.

Pennycook, Gordon, Adam Bear, Evan T. Collins, and David G. Rand (2020a), “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings.” *Management Science*, 66, 4944–4957.

Pennycook, Gordon, Jabin Binnendyk, Christie Newton, and David G Rand (2021a), "A practical guide to doing behavioral research on fake news and misinformation." *Collabra: Psychology*, 7, 25293.

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand (2018), "Prior exposure increases perceived accuracy of fake news." *Journal of Experimental Psychology. General*, 147, 1865–1880.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand (2021b), "Shifting attention to accuracy can reduce misinformation online." *Nature*, 592, 590–595.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand (2020b), "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention." *Psychological Science*, 31, 770–780.

Pennycook, Gordon and David G. Rand (2019), "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition*, 188, 39–50.

Pew Research Center (2014), "Political Polarization in the American Public."

Prior, Markus (2013), "Media and Political Polarization." *Annual Review of Political Science*, 16, 101–127.

Quattrociocchi, Walter, Antonio Scala, and Cass R. Sunstein (2016), "Echo Chambers on Facebook." *Working paper.*

Richter, Felix (2019), "The generation gap in tv consumption." *Statista.*

Rietveld, Joost (2018), "Creating and capturing value from freemium business models: A demand-side perspective." *Strategic Entrepreneurship Journal*, 12, 171–193.

Roozenbeek, Jon, Alexandra L. J. Freeman, and Sander van der Linden (2021), "How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020)." *Psychological Science*, 32, 1169–1178.

Sato, Susumu (2019), "Freemium as optimal menu pricing." *International Journal of Industrial Organization*, 63, 480–510.

Sherman, Ryland and David Waterman (2016), "The economics of online video entertainment." *Handbook on the Economics of the Internet*, 458–474.

Simmons, Steven J. (1976), "Fairness Doctrine: The Early History." *Federal Communications Bar Journal*, 29, 207.

Sims, Justin Hendrix and Paul Barrett Grant (2021), "How tech platforms fuel U.S. political polarization and what government can do about it."

Strickland, Amelia (2014), "Exploring the effects of social media use on the mental health of young adults."

Sunstein, Cass R. (2018), *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Taber, Charles S, Damon Cann, and Simona Kucsova (2009), "The motivated processing of political arguments." *Political Behavior*, 31, 137–155.

Taber, Charles S and Milton Lodge (2006), "Motivated skepticism in the evaluation of political beliefs." *American journal of political science*, 50, 755–769.

Tappin, Ben M, Gordon Pennycook, and David G Rand (2020), "Bayesian or biased? analytic thinking and political belief updating." *Cognition*, 204, 104375.

Tarski, Alfred (1955), "A lattice-theoretical fixpoint theorem and its applications." *Pacific Journal of Mathematics*, 5, 285–309.

Taylor, Sean J. and Dean Eckles (2018), "Randomized Experiments to Detect and Estimate Social Influence in Networks." In *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*, Computational Social Sciences, 289–322, Springer.

Thaler, Richard H. and Cass R. Sunstein (2009), *Nudge: Improving Decisions About Health, Wealth, and Happiness*, revised & expanded edition edition. Penguin Books, New York.

Tirole, Jean (1988), *The Theory of Industrial Organization*, 1st edition edition. The MIT Press, Cambridge, Mass.

Topkis, Donald M. (1998), *Supermodularity and Complementarity*, first edition edition. Princeton University Press.

Törnberg, Petter (2018), "Echo chambers and viral misinformation: Modeling fake news as complex contagion." *PLOS ONE*, 13.

Twenge, Jean M., Gabrielle N. Martin, and Brian H. Spitzberg (2019), "Trends in U.S. Adolescents' media use, 1976–2016: The rise of digital media, the decline of TV, and the (near) demise of print." *Psychology of Popular Media Culture*, 8, 329–345.

van der Linden, Sander, Costas Panagopoulos, and Jon Roozenbeek (2020), "You are fake news: political bias in perceptions of fake news." *Media, Culture & Society*, 42, 460–470.

Vicario, Michela Del, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi (2016), "The spreading of misinformation online." *Proceedings of the National Academy of Sciences*, 113, 554–559.

Vock, Marlene, Willemijn van Dolen, and Ko de Ruyter (2013), "Understanding Willingness to Pay for Social Network Sites." *Journal of Service Research*, 16, 311–325.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018), "The spread of true and false news online." *Science*, 359, 1146–1151.

Watts, Duncan J, David M Rothschild, and Markus Mobius (2021), "Measuring the news and its impact on democracy." *Proceedings of the National Academy of Sciences*, 118.

Wilson, Ceri and Jennifer Stock (2021), "'Social media comes with good and bad sides, doesn't it?' A balancing act of the benefits and risks of social media use by young adults with long-term conditions." *Health (London, England : 1997)*, 25, 515–534.

Yildiz, Ercan, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione (2013), "Binary opinion dynamics with stubborn agents." *ACM Transactions on Economics and Computation (TEAC)*, 1, 19.