Collective Good and Optimization in Socioeconomic Systems

Daniel E. Rigobon

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Operations Research and Financial Engineering
Adviser: Miklos Z. Rácz, K. Ronnie Sircar

May 2023

Abstract

Optimization is fundamentally grounded in perspective – one party's desired outcome may induce unintended harm on another. Such cases of misalignment between designers' incentives and collective good therefore demand attention, especially when consequences are meaningful for society. To this end, we study three settings in which individualistic optimization and social good can conflict.

First, we study how a centralized planner can modify the structure of a social or information network to reduce polarization. By formulating and analyzing a greedy approach to the planner's problem, we motivate two practical heuristics: coordinate descent and disagreement-seeking. We also introduce a setting where the population's innate opinions are adversarially chosen, which reduces to maximization of the Laplacian's spectral gap. We motivate a heuristic that adds edges spanning the cut induced by the spectral gap's eigenvector. These three heuristics are evaluated on real-world and synthetic networks. We observe that connecting disagreeing users is consistently effective, suggesting that the incentives of individuals and recommender systems may reinforce polarization.

Second, we build a model of the financial system in which banks control both their supply of liquidity, through cash holdings, and their exposures to risky interbank loans. The value of interbank loans drops when borrowing banks suffers liquidity shortages – caused by the arrival of liquidity shocks that exceeds supply. In the decentralized setting, we study banks' optimal capital allocation under pure self-interest. The second centralized setting tasks a planner with maximizing collective welfare, i.e. sum of banks' utilities. We find that the decentralized equilibrium carries higher risk of liquidity shortages. As the number of banks grows, the relative gap in welfare is of constant order. We derive capitalization requirements for which decentralized banks hold the welfare-maximizing level of liquidity, and find that systemically important banks must face the greatest losses when suffering liquidity crises – suggesting that bailouts can yield perverse incentives.

Finally, we study algorithmic fairness through the ethical frameworks of utilitarianism and John Rawls. Informally, these two theories of distributive justice measure the 'good' as either a population's sum of utility, or worst-off outcomes, respectively. We present a parameterized class of objective functions that interpolates between these two conflicting notions of the 'good'. By implementing this class of objectives on real-world datasets, we construct the tradeoff between utilitarian and Rawlsian notions of the 'good'. Empirically, we see that increasing model complexity can manifest strict improvements to both measures of the 'good'. This work suggests that model selection can be informed by a designer's preferences over the space of induced utilitarian and Rawlsian 'good'.

Acknowledgements

This thesis and all that I am is owed to the people who have shaped me throughout life. I doubt that I could ever sufficiently express my gratitude for and privilege in knowing each of them.

I am profoundly grateful to Ronnie Sircar and Miklós Rácz for all they have done over the past five years. They have enabled me to grow as a researcher, follow my ambitions, and have engaged my interests with a great deal of curiosity, patience, and compassion. I am most fortunate to have benefited from their mentorship, without which this thesis would not have been realized.

Beyond my advisors, I would like to thank several other faculty members of Princeton. I am most grateful to Sanjeev Kulkarni and Emma Hubert for serving on my dissertation committee and as reader for this thesis. Their questions and comments have helped to shape this dissertation for the better. I am also thankful for Amirali Ahmadi's feedback during my generals examination three years earlier, which have helped me to become more comfortable and confident with my work. Finally, to Matt Salganik – who I had the privilege of meeting before being accepted to this program – I owe gratitude for providing critical feedback on my research and generously sharing his time and advice.

Over the past five years, it has been a great pleasure of mine to teach. I am grateful to Ramon van Handel inspiring me with his mastery of both material and method. I am extremely thankful to Margaret Holen – for sharing with me her passion and drive for teaching, and most of all for the invaluable and unforgettable experience of teaching my first lecture. I thank Tara Zigler, Jeff Zhang, and Rajita Chandak for sharing with me the Senior Thesis Writing Group – and the privilege of paying it forward to the research community. To Joe Abbate I owe my involvement in the Prison Teaching Initiative, where I have experienced the greatest fulfillment in my day-to-day. For this I am indebted to my fellow teachers Tim Alberdingk Thijm and Ariel Bronner, along with the students at Garden State Youth Correctional Facility. I can only hope that they carry these memories with them; I know that I will.

I have been extraordinarily lucky in receiving support from the staff – especially Kim Lupinacci whose influence was felt most every day at Princeton. I thank Michael Bino for being a magician in the Terminal – but one who always reveals his tricks. I am most grateful to the dining staff of the Graduate College: Byron, Mildred, Orfenio, Roberto, and especially Luis – who one day somehow found out that I loved to eat rice and thereafter would always have a tray prepared for me at dinnertime.

I also owe a great deal to the researchers who shaped me throughout my undergraduate life.

First, to Elena Manresa, Julie Ochoa, and Abdullah Almaatouq for their patience and compassion while introducing me to their respective worlds of research. I am also most grateful to Jongwoo Lee, Neville Hogan, Alejandro Noriega-Campero, and Sandy Pentland for their kindness and investment in my growth. Without them, I would have not chosen to pursue a graduate education.

I was extremely fortunate to participate in the Fields Institute Extended Problem Solving Workshop during April of 2021, thanks to whom the COVID pandemic was colored differently. I am grateful to Mattheus Grasseli for this opportunity to participate. My thanks go out to Artur Kotlicki and Philip Schnattinger – who I am lucky to consider both friends and collaborators. I am indebted to Thibaut Duprey, whose guidance and insight has spanned far beyond our initial weeks working together. Finally, I am grateful to Tom Hurd for the contagious energy and enthusiasm he brought to our team. My memories of him will be cherished.

A great privilege of graduate school has been my fellow students. For being an extraordinary and bright cohort, I am grateful to Onat Aydinhan, Pierre Bayle, Jiarui Chu, Maycee Lin, Xiaohue Luo, Igor Silin, Yuling Yan, Tony Ye, Liz Yoo, Mengxin Yu, Louise Zhou, and Yunxiu Zhou. I would like to thank Alex Abulnaga, Simon Bilodeau, Jordan Holland, Alan Kaplan, Blake Laham, Emily Miller along with many others for being a constant and special group of friends from our early days in the Graduate College until the very end. To Felix Ackon, Sinong Geng, and Peter Tian I express my gratitude for our lovely lunches, games of basketball, and communal support while seeking out fruitful employment. For being much more than a reading group to me, I am grateful to Abraar Chaudhry, Ani Sridhar, and especially Suqi Liu – who has been a role model from early on in my graduate career, and whose presence in the office next door is sorely missed.

I am profoundly thankful for my adopted family at the Princeton Plasma Physics Laboratory: Nirbhav Chopra, Alec Griffith, Ralf Mackenbach, Eric Palmerducca, Eduardo Rodriguez Urretavizcaya, and Jacob Simmonds – without whom my graduate life would have been significantly less cheerful. I am especially grateful for their kindness, openness, sense of humor, and *excellent* cooking skills. Thank you for giving me an abundance of plasma fun facts and a community I will always treasure.

I am grateful to my longtime friends: Enzo Cerruti for the possibly-too-intellectual discussions of favorite albums, Colin Luzzi for our shared love of music, Marcus Luzzi for our many late nights playing board games, and Graham Strong for the terrifyingly closely-matched games of chess. Thanks to Jackson Graves and Tyler Ashoff for the 'study' group that remained present much beyond our courses, and to maestro Andy Tsai for constantly humbling me with his knowledge of both Go and music. I thank Daniel Mirny and Jesse Gibson for being sources of inspiration and camaraderie as

we pursue our own doctoral programs across the country. I wish you both the best – for Daniel his long-desired chalet in the French Alps, and for my longtime gym companion Jesse many more pullups. To my former roommates – Nahom Marie, Lucas Novak, and Justin Reid – I owe a great many shared evenings and meals. Although we no longer live together, I am grateful that we do not feel apart.

To Isabella – my shade on a hot day, my fire in the cold. Your energy, support, and optimism have brightened many a day. May you always be as loved as I have felt with you.

My deepest thanks go out to all of my family. First, to Ginny Isava-Toro, Francisco Lossada-Toro, and Valeria Maria Rigobon for being the first of our generation to obtain PhD's, and being inspirations along my journey. I am grateful for (an incomplete list of) those who have taught me their virtues by example: Edgar Enrique Toro of his tremendous heart and resilience; Andre Toro of drive and dedication towards one's goals; Nella Toro of cheerfulness, humor, and service in one's daily life; and Tipin of memory and remaining present in the lives of others. I must also thank Ella for the art of finding joy in life's simple pleasures – a walk through the woods on a sunny day, the smell of a leaf, or a well-timed meal.

In Ale, I am thankful to have the image of tenacity and initiative. May you continue inspiring others to follow – as I did for much of our childhood life, and beyond into the Chilean mountains. In Vero I have had a companion and life confidante – someone whose willingness to listen and help I will never doubt. I can only hope to show you the same compassion.

To my parents, I owe everything – above all my moral compass, value for family, and love of learning. They are my advisors in times of uncertainty, my celebrators during the good, and my home when all else fails. To me, their sacrifice and dedication towards their loved ones are a constant reminder of fulfillment – to live life for others. My mom has been the embodiment of love and patience – the friend and family member I strive to be. In my dad I have seen engagement, curiosity, and generosity beyond measure. I am thankful, for he is the very gentleman and scholar I hope to become. This thesis is wholly and unequivocally dedicated to them.

To my parents.

Contents

Chapter 1

Introduction

Recent decades have seen societal systems imbued with a greater degree of technological complexity. One such feature of this trend is the widespread use of optimization in decision-making. These tools often come with the promise of significant improvements in scope or quality for those who implement them, and hence are appealing in a wide variety of settings. For example – companies may use automated screening to select interviewees with reduced labor requirements, firms in the financial system may discover an allocation of their capital that reduces their exposure to risk, and social media platforms may provide better recommendations to their users. In these cases and others, there is always an inherent objective provided by designers; selection of better candidates, reducing portfolio risk, or increasing user engagement.

It is critical to note that such goals largely reflect the perspective of a single party. Namely, there is no reason to assert that designers' objectives are fundamentally aligned with the interests of other stakeholders in the system. When advancing their own agenda designers may induce harm on other parties, and depending on the particular context this externality can be significant. For illustrative purposes, let us consider the example of a company selecting candidates to fill a position. They may observe that applicants from one or more groups possess weaker qualifications. In their desire to select the 'best' from a given pool, they may systematically deprive this group from employment opportunities. Over time, this behavior could even reinforce the belief that these candidates are inherently less capable. In contrast, the very same observation could be used to hypothesize that equally-qualified applicants were adversely affected by other forces of inequality. These individuals themselves may desire a screening process guided by fair equality of opportunity, wherein their perceived disadvantages are justly contextualized.

In principle, these types of externalities emerge due to the fact that designers' share of agency is greater than their stake – the outcome is dictated by those with influence, and indifferent to the rest. In such cases, the emergent discrepancy between designers' interests and broader notions of collective good motivates a deeper understanding of potential harms. Although this misalignment can be – and often is – present without any use of optimization, the consequences in large-scale settings can clearly be more punishing. It is therefore necessary to take particular care when optimization and other algorithmic tools are used for widespread societal decision-making.

This thesis studies three distinct socioeconomic systems wherein collective good and optimization can come into conflict. Chapter 2 focuses on the emergence of opinion polarization in social networks. It formulates heuristics by which a network planner can modify the network structure in order to reduce the level of polarization. Both the formulation and findings contrast with the traditional problem of content recommendation. In Chapter 3, we model a financial system in which individual firms optimizing their portfolio – exclusively in their own best interests – causes the collective utility in the system to be reduced. In addition, regulatory interventions for ameliorating these effects are presented. Finally, Chapter 4 formulates classical machine learning problems through the lens of two distinct ethical frameworks. It relies on seeing the allocation of predictive error as a problem of distributive justice, and allows for tradeoffs between the two conflicting notions of ethical 'good'. The following sections briefly summarize the context and results of each Chapter.

1.1 Opinion Polarization in Social Networks

Social media has become prevalent in recent years, and individuals are increasingly reliant on these platforms for obtaining news and connecting with others. Paradoxically, the greater degree of connectivity in populations has coincided with stronger polarization of opinions. It is believed that social media plays an important role – users can change their opinions depending on the information and content they are exposed to. From the perspective of social media platforms, content mediation is often driven by so-called recommender systems, which aim to expose a user to pieces of content that are similar to those they have liked. In users, the well-studied phenomenon of confirmation bias leads them to seek out content that aligns with their current opinions. It has been seen that these two forces contribute to increasing polarization (Chitra and Musco, 2020; Bhalla et al., 2021).

Chapter 2 aims to assess how the structure of social networks can contribute to opinion polarization. In particular, it asserts that *how* users are connected to each other is critical for understanding the level of polarization. We study the effects of both a fixed social network, and small modifications

made to its edges in the interest of reducing polarization. It is important to note that we consider polarization to be an undesirable feature for a population – in which case the incentives of both individuals and social media platforms will fail to bring opinions closer to a consensus.

This chapter uses tools from spectral graph theory and models of opinion dynamics. In part, it contains work published in Racz and Rigobon (2023).

Main Results There are three branches of theoretical contributions to highlight, along with several simulations.

First, we will see how polarization is tied to structural properties of social networks. In particular, it is shown that the presence of bottlenecks in the graph can contribute to greater levels of polarization. Therefore, echo chambers and tightly-knit community structures can be liabilities – uniformly connected networks are most conducive to low-polarization outcomes.

Second, we present an optimization problem for a planner who can perturb the network's structure in order to reduce polarization. In stark contrast to this problem, traditional recommendation systems will often operate with the goal of increasing user engagement or satisfaction. We will see that the largest reductions in polarization can be obtained when disagreeing individuals are connected, which suggests that neither users nor platforms would create these connections when operating selfishly.

Third, we study a different setting for the planner wherein the population's opinions are unknown. The resulting problem of minimizing worst-case polarization reduces to a fundamental problem in spectral graph theory – the maximization of a graph's spectral gap. Colloquially, this planner will aim to reduce the presence of any strong bottlenecks in the graph structure. We will see that adding edges between different communities is provably effective for doing so.

In practice, the planner's optimization problem is challenging to solve efficiently. However, the previous theoretical results can be used to inform principled heuristics for reducing polarization. In experimental results, we test these heuristics on several real and synthetic social networks. We will see that in many cases, polarization can be significantly reduced with a small number of edge modifications. This suggests that in some cases polarization may be a fragile phenomenon of current social network structures.

1.2 Risks in Formation of Interbank Lending Networks

History has shown that financial systems can be at risk of distress spreading throughout firms. For example, in the crisis of 2008, it was believed that the failure of so-called 'systemically important' institutions would lead to widespread losses within both the financial sector and beyond. The result was a substantial amount of government aid provided to assist distressed banks. For these risks to have been present, it would be necessary for firms to be connected in some manner – either directly or indirectly. While these linkages may be beneficial during normal times, they can be liabilities when shocks occur, and ultimately facilitate the phenomenon of financial contagion. Nonetheless, it is conceivable that firms form these linkages in a strategic manner – in their own best interest.

Chapter 3 presents a model of a simple financial system in which banks concurrently make decisions to lend capital to each other. They do so following a core paradigm of financial risk management – portfolio optimization. In doing so, banks must also control their own exposure to idiosyncratic shocks. When a bank suffers an overwhelmingly large shock, both itself and its creditors will suffer losses – allowing for instances of localized distress to have more global effects. We characterize inefficiencies in this system and study how regulation might be used to improve the system's overall welfare.

This work leverages techniques in financial mathematics, stochastic analysis, and optimal control. It contains results published in a preprint by Rigobon and Sircar (2022).

Main Results We begin by building a novel model of the banking system in which banks control both their supply of liquidity, through cash holdings, and their exposures to risky interbank loans. The value of interbank loans jumps when banks suffer liquidity shortages, which can be caused by the arrival of large enough liquidity shocks. This results in a simple game-theoretic setting, where each firms' stochastic payoffs depend on the others' level of liquidity risk.

In two distinct settings, we compute the unique optimal allocations of capital. In the first, banks behave with pure self-interest and seek only to maximize their own utility – termed the 'decentralized' equilibrium. However, this equilibrium suffers from an important externality – banks are not punished when their liquidity shortages cause harm to creditors. The second 'centralized' equilibrium follows from a single planner who determines how all capital in the system is to be allocated. The planner captures the maximal value that can be obtained in the system, but the centralized equilibrium is unstable from the perspective of individual banks. Under technical conditions, existence and uniqueness of both equilibria are shown.

By comparing the two equilibrium allocations, we see that the decentralized equilibrium is more likely to suffer from liquidity shortfalls that induce systemic losses. In addition, we can quantify the value lost due to banks' greedy behavior. This notion of the 'price of anarchy' is found to be of constant order in the size of the financial system, suggesting that larger systems are not relatively more inefficient. We do, however, observe that in the centralized equilibrium, the likelihood of a bank having insufficient liquidity shrinks with the system size. This touches on the crucial externality in the model, wherein individualistic behavior reduces collective value.

Finally, we show how regulation – in the form of capitalization requirements – can force banks to closely replicate the centralized equilibrium. In doing so, we find that banks with a large number of counterparties must face the greatest losses when they suffer liquidity shortages – ensuring that they are incentivized to avoid such crises.

1.3 Tradeoffs in Algorithmic Fairness

As algorithmic decision-making has become more impactful, concerns have been raised regarding the ethics of their effects. It is especially important to address these concerns when decisions can have significant and meaningful consequences on humans. Classic examples in criminal justice and hiring practice have shown that firms can exhibit discriminatory behavior – which can be accentuated by the use of machine learning. Motivated by these observations, researchers have sought to answer several questions, among which two are: 1) how does one design a fair algorithm? and 2) what are the costs associated with addressing unfairness? Rooted in any answer to these questions must be an underlying definition of fairness itself – many existing works use some kind of egalitarianism. However, there are a large number of possible ethical frameworks for defining and addressing fairness.

In Chapter 4, we study how two conflicting frameworks for distributive justice can be used to inform algorithm design. This work proceeds from the fundamental assumption that the optimal allocation of 'loss' (.e. predictive error of a model) can be viewed as a problem of distributive justice. In particular, we study the tradeoff between utilitarian and Rawlsian models. It is valuable to consider this tradeoff because each of these theories of justice can exhibit its own unique shortcomings, which can be addressed by the other – utilitarianism suffers from an indifference to inequality, and Rawlsian design fails to account for the majority of the population. A core feature of this work is the ability to consider 'fairness' through a model designer's preferences over bundles of utilitarian and Rawlsian 'good'.

This chapter combines ideas from a broad set of disciplines, including ethics, welfare economics,

computer science, and statistics. It contains work published in Rigobon (2023).

Main Results The fundamental contribution of this chapter is a continuum of objective functions for learning problems that allows one to interpolate between two perspectives: utilitarian and Rawlsian. We assert that a utilitarian designer considers only the minimization of average-case loss, whereas the Rawlsian designer optimizes for worst-case loss. The minimization of each objective function along the proposed continuum yields a different 'optimal' model – which is tied to the minimization of a representative individual's expected disutility. In this interpretation, the continuum of objectives is parameterized by a 'risk aversion' parameter, such that the utilitarian model corresponds to risk neutrality, and the Rawlsian model to infinite risk aversion. These results are shown formally through convergence of the sequence of minimizers. By projecting each optimal model onto the space of utilitarian and Rawlsian 'good', we can obtain a frontier that illustrates the tradeoff between the two 'goods'. In practice, this frontier can be used to understand the marginal rate of transformation between these two conflicting notions of the good. Through this perspective, it is possible to incorporate preferences for fairness – not as a constraint, but as a fundamental part of the objective.

In experiments, we compute the tradeoff for several common datasets. In addition to observing the shape of the tradeoff itself, we study how increasing the predictive expressibility of a model can manifest improvements to both average-case (resp. utilitarian) and worst-case outcomes (resp. Rawlsian good). In doing so, we argue that fairness-as-a-constraint paradigms may sacrifice large reductions in worst-case error in the name of minimal improvements to average-case error.

Finally, we present several open problems – both theoretical and empirical – that would help further explore the features of this tradeoff, and how algorithmic fairness can be more broadly guided by ethical principles.

Chapter 2

Opinion Polarization in Social Networks

2.1 Introduction

In recent years there has been a substantial increase in sociopolitical polarization – it is clear that our society does not agree on issues in politics, science, healthcare, and beyond. Counter-intuitively, this has been accompanied by the growth of social media platforms; individuals are connecting with others and sharing information more than ever before. How is it that "bringing the world closer together"[1] resulted in our opinions drifting further apart?

This phenomenon is a byproduct of the structure of our social networks; a greater number of connections does not necessarily reflect a closeness to consensus. It is possible for the proliferation of social media to reduce one's exposure to other opinions, and thereby entrench them in a community of like-minded users. This feature is known as an "echo chamber," and has been found to emerge through the incentives of recommender systems rewiring the network (Chitra and Musco, 2020). Furthermore, confirmation bias and structural similarity have been found to contribute to increases in polarization as the structure of the network evolves (Bhalla et al., 2021; Santos et al., 2021). Therefore, *how* the population is connected – as opposed to how connected the population *is* – may be most important to the emergence of polarization.

In this chapter, we seek an understanding of how a network planner can reduce polarization by changing the structure of a population's social or information network. To that end, we present a model of *budgeted network perturbation*, where the planner is given a small budget with which to modify the structure of a given network. We study the planner's problem in two different settings,

[1] The original mission statement of Facebook.

and evaluate simple heuristics on both real-world and synthetic networks.

There has been a significant research effort towards reducing polarization in networks (Chen et al., 2018; Garimella et al., 2017; Haddadan et al., 2021; Matakos et al., 2017; Rahaman and Hosein, 2021). In contrast to both Matakos et al. (2017) and Rahaman and Hosein (2021), we hold fixed the population's opinions – while allowing the network structure to be modified. This work differs from Garimella et al. (2017) and Haddadan et al. (2021) in both our use of a distinct measure of polarization and incorporating opinion dynamics. Finally, we improve upon the closely related work of Chen et al. (2018) through a more detailed theoretical analysis of edge effects, consideration of weighted networks, and study of larger datasets.

A very similar contribution to our own is recent work by Zhu et al. (2021), where the authors present a variation of the problem studied in Musco et al. (2018). Both these studies aim to minimize the sum of polarization and disagreement by changing the network structure, but Zhu et al. (2021) impose a budget that ensures only a small number of edges can be changed. These authors use a similar budget constraint to our own, but their polarization-disagreement index varies greatly with the edge density of the graph. Although it is convenient for analysis and computation, their index is inadequate for capturing the dynamics of polarization alone. Nonetheless, we believe the formulation in this chapter and Zhu et al. (2021) to be practical. The network structure is not assumed to be completely malleable, but small changes are permitted. For instance, while social media platforms such as Facebook or Twitter cannot dictate who an individual chooses to 'friend' or 'follow', these platforms can curate an individual's feed to change one's relative exposure levels to certain content. This process perturbs the structure of external influence on an individual, so that it differs from their endogenously created network of 'friends' or 'follows'. If, instead, any of these platforms suddenly decided to completely rewire their social networks, users may be upset.

It is then natural to consider the following questions: how does the network planner decide to allocate their budget? How much of an impact can be made? How large of a budget is needed to achieve a significant reduction in polarization?

We begin by first establishing a relationship between structural properties of a social network and its level of polarization. We find that both the degree profiles and the strength of information bottlenecks – quantified by the well-known Cheeger constant in spectral graph theory – are closely tied to polarization. This result naturally captures the intuition and dangers of echo chambers in real-world networks.

Next, we focus on the formulation and analysis of two settings for network optimization. In the first, the planner has full information of the population's opinions. We provide theoretical motivation

for two heuristics: coordinate descent and a stepwise disagreement-seeking algorithm. The former is standard in optimization, while the latter is the antithesis of confirmation bias. Existing research has shown that addition of edges between like-minded individuals contributes to increasing polarization (Bhalla et al., 2021). Moreover, according to Bindel et al. (2015) it is 'costly' for individuals to be connected to others who disagree with them, and recommender systems can be designed to minimize disagreement (Chitra and Musco, 2020). Therefore, the incentives of both individuals and social media platforms may naturally lead to polarization growing over time. In contrast, we show that a simple disagreement-seeking approach taken by the planner leads to substantial reductions in polarization. This result is closely tied to our choice of the opinion dynamics model. In this work, interactions between individuals are always *attractive* – bringing opinions closer together. In reality, this is not the case (see Bail et al. (2018); Balietti et al. (2021)). This simplification, however, will facilitate theoretical results – which we believe can be leveraged for partial understanding of polarization-reduction strategies in a richer class of models.

This chapter also presents a novel setting for the network planner, wherein the population's opinions are chosen adversarially. In several papers from the literature (see, for instance Chen and Rácz (2022); Gaitonde et al. (2020); Matakos et al. (2017); Rahaman and Hosein (2021)), an adversary is able to change the individuals' opinions – seeking to maximize polarization. The setting we study represents a planner whose network design must be *robust* to the adversary's disruption. We show that this setting for the planner's problem is intimately related to maximizing the spectral gap of the graph's Laplacian, which is a well-studied problem (Donetti et al., 2006; Wang and Van Mieghem, 2010; Watanabe and Masuda, 2010). We provide theoretical guarantees for a heuristic that connects vertices on opposite sides of the cut corresponding to the spectral gap.

We then evaluate several natural heuristics on real-world and synthetic networks. There are significant reductions in polarization for networks with strong initial community structures. Furthermore, we study how the spectral gap and homophily are affected by the planner's modifications. We find that the largest reductions in polarization are accompanied by reductions in homophily. In many cases, however, one of our heuristics effectively reduces polarization with little effect on homophily. We also observe that two heuristics lead to vertices with extreme opinions becoming more central in the graph structure. In many of the networks studied, a small budget yields substantial reductions in polarization.

The chapter is organized as follows. Section 2.1.1 provides a detailed review of recent and related work. Next, Section 2.2 introduces relevant notation, definitions, and preliminaries. Section 2.3 provides theoretical ground for three heuristics, which are described and evaluated on several networks

in Section 2.4. Finally, Section 2.5 concludes and discusses potential directions for future work.

2.1.1 Relevant Literature

The papers most similar to our work are recent studies by Chen et al. (2018), Gaitonde et al. (2020), Chitra and Musco (2020), and Zhu et al. (2021). Gaitonde et al. (2020) motivates our adversarial disruption of the population's opinions, while both Chen et al. (2018) and Zhu et al. (2021) aim to modify a social network's structure by adding a small number of edges. Chitra and Musco (2020) impose a constraint on the edge weight modified – but not the number of edges. In particular, they focus on changing a large number of edges by a small amount, whereas we seek to do the converse. Our work differs from Chen et al. (2018) through greater emphasis on theory and generalization to weighted graphs. The objective function in Zhu et al. (2021) fundamentally differs from our own, and represents a different problem faced by the network planner.

In addition, this work is broadly tied to the literature on opinion dynamics, perturbation of network structures, and influencing polarization. Relevant studies in each of these areas are discussed in the following.

Opinion Dynamics

The study of consensus-forming begins with the seminal work of DeGroot (1974), where under weak conditions on the social network, the opinions eventually converge to a perfect consensus. This model was expanded by Friedkin and Johnsen (1990) (and more recently by Conjeaud et al. (2022)), so that the long-term opinions are heterogeneous. Because of this feature and its simplicity, the Friedkin-Johnsen (FJ) model has appeared in several recent studies on opinion polarization and disagreement – see for instance, Matakos et al. (2017); Musco et al. (2018); Chen et al. (2018); Chitra and Musco (2020); Gaitonde et al. (2020); Zhu et al. (2021); Chen and Rácz (2022). We will also use the same FJ model. Not only is it standard in the literature, but it is mathematically convenient for analysis. There are also rich areas of work which justify and extend the FJ model. For instance, Bindel et al. (2015) show that the expressed opinions of this model correspond to the Nash equilibrium of a cost-minimizing game between individuals.

There are a few notable extensions to the FJ model, in which individuals have more complex behavior. For example, a recent survey by Biondi et al. (2022) presents several generalizations and (relevantly) assesses if polarization can occur in each. A fundamental feature of the FJ model is that individuals are always drawn toward the opinions of their neighbors – but experimental

evidence of this feature is inconclusive and contextual (Bail et al., 2018; Balietti et al., 2021). Motivated by this observation, new models have been developed in which individuals have bounded confidence (Hegselmann et al., 2002) or even experience repulsion (Rahaman and Hosein, 2021; Cornacchia et al., 2020). It is also possible to incorporate geometric structures into the dynamics, such as recent work by Hązła et al. (2019) and Gaitonde et al. (2021). Finally, we note that there are several related studies within the controls literature, which focus on consensus dynamics on a network, for instance, when agents have antagonistic dynamics (Altafini, 2012), or are stubborn (Mao et al., 2018) – see Qin et al. (2016) for a more complete survey.

Optimizing Network Structures

This chapter formulates an optimization problem over network structures, aiming to reduce a particular definition of polarization. There are several related works in the literature. For example, Musco et al. (2018) allows unconstrained rewiring of the social network to reduce the *polarization-disagreement index*, which is defined as the sum of polarization and disagreement. A recent paper of Zhu et al. (2021) optimizes the same index via addition of a limited number of edges. This index is analytically and computationally convenient because of its monotonicity and convexity, but it is highly sensitive to the edge density of the graph.[2] We instead focus exclusively on minimizing polarization, which is shown to be neither convex nor monotone in Section 2.3.2. However, this paper restricts edge modifications similarly to Zhu et al. (2021).

A more closely related work by Chen et al. (2018) presents several definitions of 'conflict' in social networks, and studies how they can be minimized through iterative perturbations to the graph. One such measure of conflict equals polarization. We expand on the authors' work by providing a detailed theoretical analysis of edge perturbations on polarization, generalizing the analysis to weighted graphs, and conducting simulations on larger real-world and synthetic networks.

The aforementioned papers share with our work a definition of polarization. However, it is possible to optimize for other notions of 'cohesiveness' or 'consensus'. For instance, Garimella et al. (2017) and Haddadan et al. (2021) both present measures of polarization based on random walks, and propose algorithms for reducing it via edge addition. The greatest similarity between their work and ours lies in the use of a greedy, stepwise approach to a combinatorial optimization problem. However, the authors' definitions of polarization do not directly incorporate opinion dynamics.[3]

[2]The polarization-disagreement index consists of adding polarization, which is on the order of n (the number of vertices), and disagreement, which is of order m (the number of edges). Therefore this index is dominated by disagreement for dense graphs (specifically, if $m \gg n$).

[3]We note that the Friedkin-Johnsen model has a random walk interpretation of the long-term opinions, see Gionis et al. (2013).

Moreover, in Haddadan et al. (2021), nodes represent webpages, not individuals.

Another definition of cohesiveness, which does not depend on any node opinions or labels, is the spectral gap of a graph. The spectral gap controls the synchronizability of dynamical systems and mixing times of Markov chains (Donetti et al., 2006), and therefore its maximization is of great interest. For instance, Watanabe and Masuda (2010) seek to increase the spectral gap by removing nodes. Unlike these authors, we focus on changes to a graph's edges. More relevantly, Wang and Van Mieghem (2010) study how the algebraic connectivity (i.e., spectral gap) can be increased by adding edges. The authors present two strategies for doing so, one of which is derived from the eigenvector corresponding to the spectral gap. We show that the adversarial setting of the planner's problem is closely related to their work, and provide bounds on polarization using this eigenvector-based strategy.

Natural Network Dynamics

A different branch of research aims to understand how polarization is shaped by *rewiring dynamics* in the network. For instance, a recent paper by Bhalla et al. (2021) studies how individuals' local rewiring rules can lead to higher polarization. The authors conclude that confirmation bias and friend-of-friend behavior are critical for this result. However, their theoretical results focus on the polarization-disagreement index. Moreover, we derive an improved upper bound for polarization in Section 2.3.1. A similar paper by Santos et al. (2021) shows that allowing individuals to rewire according to structural similarity leads to polarization, although the authors use a distinct model of opinion dynamics.

It is also possible to study the dynamics driven by a network administrator. Chitra and Musco (2020) present a setting in which a network administrator rewires the network over time by providing 'recommendations' to users based on minimizing disagreement. They show that without a regularization term in the optimization problem, the administrator greatly increases polarization. The authors' result contrasts with one of the main findings of this chapter, namely that connecting disagreeing individuals is effective for reducing polarization.

Optimizing Opinion Profiles

While less relevant to this work, a complimentary line of work assumes that the network structure remains fixed, but the innate opinions are subject to change. For instance, Gionis et al. (2013) establish NP-Hardness of an opinion maximization problem, in which an administrator takes over a small set of individuals and sets their opinions to the largest possible value. Papers by Matakos

et al. (2017) (resp. Matakos et al. (2020)) seek to minimize polarization (resp. maximize diversity, i.e., disagreement) by choosing a small subset of individuals to have neutral opinions. Finally, the work of Rahaman and Hosein (2021) aims to minimize polarization in an extension of the FJ model, but by shifting each individuals opinion by a small amount.

These studies have generally taken the perspective of a benevolent network planner. It is also possible to consider the perspective of an adversary, who takes over a small number of individuals and seeks to maximize polarization or disagreement (Chen and Rácz, 2022). A more powerful adversary in Gaitonde et al. (2020) chooses the opinions of the entire population to the same end. In particular, Gaitonde et al. (2020) present a problem of *defending* the network from this adversary by making some opinions more resistant to change. We will consider a similar setting, but where the network is defended by altering its structure instead. Nonetheless, the adversary faced is modeled on their work.

2.2 Model

An undirected graph $\mathcal{G}(V, E, W)$ is defined by a set of vertices V given by $[n] := \{1, \ldots, n\}$, a set of edges $E \subset V \times V$ consisting of unordered pairs of vertices, and weight matrix $W \in [0, \bar{w}]^{n \times n}$. W is assumed to be a symmetric matrix of non-negative edge weights such that $w_{ij} > 0$ if and only if $(i, j) \in E$, and $\bar{w} < \infty$ indicates the maximum possible edge weight. For a graph $\mathcal{G}$, its degree matrix D is diagonal, and satisfies $D_{ii} = d_i$, where $d_i = \sum_j w_{ij}$ is the (weighted) degree of vertex i. Let $L = D - W$ denote the combinatorial graph Laplacian, and $\mathcal{L} = D^{-1/2} L D^{-1/2}$ denote the normalized Laplacian. We write $N(i) := \{j \in [n] : (i, j) \in E\}$ for the neighbors of vertex i. For a matrix $A \in \mathbb{R}^{n \times n}$, we will write $\lambda_n(A) \geq \ldots \geq \lambda_2(A) \geq \lambda_1(A)$ to denote its eigenvalues in descending order.

Vertices are given *innate opinions* $\mathbf{s} \in [0, 1]^n$, which represent a continuum between two extreme positions on an issue. For instance, an individual who is totally in favor of strict firearm laws may have an opinion of 0, whereas one extremely against any such regulations would have an opinion of 1. The population's opinions evolve over time, beginning from the innate opinions $\mathbf{s}$. The evolution of opinions follows the dynamics of Friedkin and Johnsen (1990) (see below), and the opinions converge to a fixed point – denoted $\mathbf{z}$ and called the *expressed opinions* of the population. We are interested in modifications to the underlying graph $\mathcal{G}$, and therefore take the innate opinions $\mathbf{s}$ to be fixed. Consequently, we write $\mathbf{z}$ and $\mathbf{z}'$ for the expressed opinions corresponding to the social networks $\mathcal{G}$ and $\mathcal{G}'$, respectively. Occasionally, to emphasize the underlying graph $\mathcal{G}$, we will write $\mathbf{z}_{\mathcal{G}}$.

2.2.1 Opinion Dynamics

In the seminal model of DeGroot (1974), the population's expressed opinions converge to a perfect consensus under weak conditions. A notable extension of the DeGroot model is by Friedkin and Johnsen (1990), whose model preserves long-term heterogeneity of opinions. In particular, $\mathbf{z} = c\vec{\mathbf{1}}$ if and only if $\mathbf{s} = c\vec{\mathbf{1}}$. This model is convenient for analysis because the expressed opinions can be written explicitly. Furthermore, several recent works in the literature have leveraged this opinion dynamics model – see Section 2.1.1 for more detail.

The Friedkin-Johnsen (FJ) opinion dynamics model is specified by the discrete-time mapping $\mathbf{s}(t) \to \mathbf{s}(t+1)$ as follows. We initialize $\mathbf{s}(0) = \mathbf{s}$, and iterate

$$s_i(t+1) = \frac{s_i(0) + \sum_{j \in N(i)} w_{ij} s_j(t)}{1 + \sum_{j \in N(i)} w_{ij}},$$

where w_{ij} is the weight associated with edge (i, j), and is non-zero if and only if $j \in N(i)$. The expressed opinions $\mathbf{z}$ are the fixed point of this mapping, given by

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

where I denotes the $n \times n$ identity matrix. Notice that $I + L \succcurlyeq I$ is necessarily invertible. Thus, there exist unique expressed opinions $\mathbf{z}$ for any given $\mathcal{G}$ and s. Moreover, since the eigenvalues of $(I+L)^{-1}$ are no greater than 1, the expressed opinions of the FJ dynamics are a contraction of the innate opinions. This observation also follows from the fact that the FJ model is purely *attractive* – opinions of connected individuals are always drawn to each other over time. One of the heuristics we develop will depend on this feature of the dynamics. However, exposure to substantially differing opinions in the real-world may yield no effect, or even strengthen one's original position. In Section 2.5 we discuss how our results might be leveraged for such a class of richer opinion dynamics models, and relevant directions for future work.

2.2.2 Polarization and Disagreement

In practice, a perfect consensus is rare; therefore, we seek to understand "closeness" to consensus. Accordingly, we define *polarization* to be proportional to the variance of the expressed opinions. Large polarization indicates that the population is far from achieving a consensus, and vice-versa. Formally, we define:

Definition 2.1 (Polarization)**.** *Given a vector of opinions* $\mathbf{x} = (x_1, \ldots, x_n)$ *and the mean of its*

entries $\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$*, the* polarization *of* $\mathbf{x}$ *is*

$$P(\mathbf{x}) := \sum_{i=1}^{n} (x_i - \overline{x})^2 = \|\widetilde{\mathbf{x}}\|^2,$$

where $\widetilde{\mathbf{x}} := \mathbf{x} - \overline{x}\vec{\mathbf{1}}$ *are the mean-centered opinions.*

In particular, $P(\mathbf{z})$ is *expressed* polarization, and $P(\mathbf{s})$ is *innate* polarization.

It is also useful to define *disagreement*, which captures distance from consensus on a local scale. Intuitively, if two vertices have very distinct opinions, then their disagreement is large.

Definition 2.2 (Disagreement)**.** *For any vector of opinions* $\mathbf{x} = (x_1 \dots x_n)$*, the disagreement between vertices* i *and* j *is given by:*

$$D_{ij}(\mathbf{x}) := (x_i - x_j)^2.$$

Again, between vertices i and j, $D_{ij}(\mathbf{z})$ is the *expressed* disagreement, while $D_{ij}(\mathbf{s})$ is the *innate* disagreement. The two quantities above have been studied in several recent papers on social and information networks; see Matakos et al. (2017); Musco et al. (2018); Chen et al. (2018); Chitra and Musco (2020); Gaitonde et al. (2020); Zhu et al. (2021); Bhalla et al. (2021); Rahaman and Hosein (2021); Santos et al. (2021); Chen and Rácz (2022) and references therein.

2.3 Theoretical Results

We now present several theoretical results on polarization. We study how its magnitude depends on structural properties of the graph, and how it can vary as a planner modifies the edges.

2.3.1 Opinion Contraction and Polarization

We are primarily concerned with polarization of expressed opinions, $P(\mathbf{z})$. However, the relationship between expressed and innate polarization depends on $\mathcal{G}$. Since the opinion dynamics model performs a contraction on the opinions, it follows that $P(\mathbf{z}) \leq P(\mathbf{s})$. In fact, more is true: the contraction ratio is controlled by the degrees and structural properties of $\mathcal{G}$.

To present this result, we must introduce some notation. For any two disjoint subsets of vertices B_1 and B_2, let $E(B_1, B_2)$ denote the set of edges with one incident vertex in B_1 and the other in B_2. We define the conductance of a nonempty subset of vertices X as

$$h_{\mathcal{G}}(X) := \frac{\sum_{(i,j)\in E(X,X^C)} w_{ij}}{\min\{\sum_{v\in X} d_v, \sum_{u\in X^C} d_u\}}.$$

The isoperimetric number (also known as the Cheeger constant) of a graph $\mathcal{G}$ is given by

$$h_{\mathcal{G}} := \min_{X\subset V, 0<|X|<|V|} h_{\mathcal{G}}(X), \tag{2.1}$$

as in Chung (1997), and will appear in the results. Note that $h_{\mathcal{G}} \leq 1$, since $h_{\mathcal{G}}(X) = 1$ when X consists of a single vertex. Furthermore, $h_{\mathcal{G}} = 0$ if and only if $\mathcal{G}$ is disconnected. The isoperimetric number of a graph is an indication of the presence of bottlenecks – it is small when there exists a large set of vertices that is sparsely connected to the remainder of the graph.

We now arrive at a first result on the contraction properties of the FJ model on polarization.

Proposition 2.1. *Let $d_{\min}$ and $d_{\max}$ be the minimum and maximum weighted degrees in $\mathcal{G}$, and let $h_{\mathcal{G}}$ be its isoperimetric number. Then,*

$$\frac{P(\mathbf{s})}{\left(1 + (2d_{\max}) \wedge (\bar{w}n)\right)^2} \leq P(\mathbf{z}) \leq \frac{P(\mathbf{s})}{\left(1 + \frac{1}{2}d_{\min}h_{\mathcal{G}}^2\right)^2}.$$

Proposition 2.1 quantifies the effects of the FJ model on polarization. In particular, if $\mathcal{G}$ has strong expander properties (i.e., $h_{\mathcal{G}}$ is large), then we expect the expressed polarization to be small, relative to the innate polarization. The proof of this result can be found in Appendix 2.A, and follows from simple eigenvalue bounds and a version of Cheeger's inequality.

This result provides a tighter upper bound on polarization than that of Bhalla et al. (2021). The tightening is achieved by observing that the mean-centered innate opinions $\widetilde{\mathbf{s}}$ are orthogonal to the eigenvector of $(I+L)^{-2}$ that has corresponding eigenvalue 1. In addition, we can use Proposition 2.1 to show that the complete graph K_n, with all edge weights equal to the maximal $\bar{w}$, is a global minimum for polarization.

Corollary 2.2. *Fix innate opinions $\mathbf{s}$, and let $\mathcal{G}$ be any graph on n vertices with maximal edge weight $\bar{w}$. Let $\mathbf{z}_{K_n}$ and $\mathbf{z}_{\mathcal{G}}$ denote the expressed opinions on K_n and $\mathcal{G}$, respectively. Then,*

$$P(\mathbf{z}_{K_n}) \leq P(\mathbf{z}_{\mathcal{G}}).$$

Moreover, $P(\mathbf{z}_{K_n}) = \frac{P(\mathbf{s})}{(1+\bar{w}n)^2}$.

The key observation in the proof of Corollary 2.2 is that all non-zero eigenvalues of $L_{K_n} = \bar{w}\left(nI - \vec{\mathbf{1}}\vec{\mathbf{1}}^T\right)$ (the Laplacian of the complete graph) are equal to $\bar{w}n$. Therefore, for any $\mathcal{G}$, the

value of polarization on K_n achieves with equality the smallest lower bound from Proposition 2.1. This result also provides a useful reference point for indicating the planner's closeness to global optimality.

2.3.2 Given Opinions

We now turn to studying how the planner can decrease polarization by modifying the graph.

In a first setting, we assume that the innate opinions are known. If the planner can change (by adding or removing) the edge weight between at most k pairs of vertices, what is the least polarization they can achieve? Formally, given a graph $\mathcal{G}$ with innate opinions $\mathbf{s}$, and integer budget $k > 0$, we wish to solve

$$\begin{aligned} &\min_{\mathcal{G}'} P(\mathbf{z}') \\ &\text{s.t. } ||W - W'||_0 \leq 2k, \end{aligned} \tag{2.2}$$

where the expressed opinions $\mathbf{z}'$ correspond to $\mathcal{G}'$, which must also be an undirected graph with maximal edge weight $\bar{w}$. The factor of two in the constraint of (2.2) follows from our assumption of undirected graphs. The constraint naturally captures the assumption that it is costly for the planner to modify an edge, but upon committing to doing so, they may freely change the edge weight.[4]

Problem (2.2) is challenging to solve efficiently since it is non-convex. Beyond the fact that the ℓ_0 constraint gives a non-convex feasible set, the objective function (over valid Laplacian matrices) is also not convex – see Figure 2.1 for a small example. Therefore, relaxing the ℓ_0 constraint to ℓ_1 will still yield a non-convex optimization problem. Instead of seeking an optimal set of k edges to add, we propose a greedy stepwise approach where the weight of k edges are saturated iteratively, one at a time. This simpler setting is tractable for analysis.

It seems intuitive that adding edge weight to $\mathcal{G}$ promotes the flow of information, and thereby reduces polarization. However, this is not the case in general. We will see that for most non-saturated edges, there exists a value of the innate opinions for which the addition of weight to that edge will increase polarization. The exact expression for the change in polarization when adding edge weight is given in the following.

[4]In principle, the constraint may bound the absolute difference in edge weights (ℓ_1 norm). This is an entirely different problem, (more similar to Chitra and Musco (2020)) but an interesting direction for future work. We believe that with an ℓ_1 constraint, the planner would distribute its edge weight to maximize the minimal marginal return of polarization with respect to edge weight. We will also see that relaxing the ℓ_0 constraint to ℓ_1 is insufficient for obtaining a convex optimization problem.

(a) Graph $\mathcal{G}_1$ (b) Graph $\mathcal{G}_2$

Figure 2.1: A simple example of the non-convex objective function. With innate opinions $\mathbf{s} = [0, 0.4, 1]$, it can be seen that $P\left(\frac{1}{2}[L_1 + L_2]\right) > \frac{1}{2}[P(L_1) + P(L_2)]$. (Note the abuse of notation to illustrate $P(\cdot)$'s dependence on only the Laplacian.) In this particular example, the addition of any amount of weight to edge $(1, 3)$ *increases* polarization.

Lemma 2.3. *Let $\mathcal{G}(V, E)$ be an undirected graph yielding expressed opinions $\mathbf{z}$, and (i, j) be a pair of vertices with non-maximal weight, that is, $w_{ij} < \bar{w}$. Let $\mathbf{v}_{ij} := \mathbf{e}_i - \mathbf{e}_j$. For $\delta \in (0, \bar{w} - w_{ij}]$, we construct $\mathcal{G}^+(V, E^+, W^+)$ according to $w^+_{ij} = w_{ij} + \delta$, and $E^+ = \{(i, j) : w^+_{ij} > 0\}$. If the expressed opinions on $\mathcal{G}^+$ are given by $\mathbf{z}^+ := (I + L^+)^{-1}\mathbf{s}$, then*

$$P(\mathbf{z}) - P(\mathbf{z}^+) = D_{ij}(\mathbf{z}) \left[\frac{2\delta \widetilde{\mathbf{z}}^T (I + L)^{-1} \mathbf{v}_{ij}}{\widetilde{\mathbf{z}}^T \mathbf{v}_{ij} \left(1 + \delta \mathbf{v}_{ij}^T (I + L)^{-1} \mathbf{v}_{ij}\right)} - \frac{\delta^2 \mathbf{v}_{ij}^T (I + L)^{-2} \mathbf{v}_{ij}}{\left(1 + \delta \mathbf{v}_{ij}^T (I + L)^{-1} \mathbf{v}_{ij}\right)^2} \right]. \tag{2.3}$$

The proof of this result can be found in Appendix 2.A. To discuss this result in more detail, it is useful to define the following.

Definition 2.3 ($\partial_{w_{ij}} P(L)$)**.** *Fix some innate opinions $\mathbf{s}$. Let $\mathbf{z}_L$ denote the resulting expressed opinions when the underlying graph $\mathcal{G}$ has Laplacian L. We write:*

$$\partial_{w_{ij}} P(L) = \lim_{t \to 0^+} \frac{P(\mathbf{z}_{L + tL_{ij}}) - P(\mathbf{z}_L)}{t} \tag{2.4}$$

where $L_{ij} = \mathbf{v}_{ij}\mathbf{v}_{ij}^T$.

This definition allows us to analyze the first-order effects of edge modifications on polarization. Notice that even if a graph were unweighted, we can define this derivative for its equivalent *weighted* graph, where the weight of each existing edge equals one. In the following proposition, we derive a closed form expression for these partial derivatives.

Proposition 2.4. *For fixed innate opinions $\mathbf{s}$, we have*

$$\begin{aligned} \partial_{w_{ij}} P(L) &= -2\widetilde{\mathbf{s}}^T (I + L)^{-2} L_{ij} (I + L)^{-1} \widetilde{\mathbf{s}} \\ &= -2\widetilde{\mathbf{z}}^T (I + L)^{-1} L_{ij} \widetilde{\mathbf{z}}. \end{aligned}$$

This result allows us to re-write (2.3) in Lemma 2.3 as:

$$P(\mathbf{z}) - P(\mathbf{z}^+) = \frac{-\delta \partial_{w_{ij}} P(L)}{1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} - \frac{\delta^2 \mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}}{\left(1 + \delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}\right)^2} (z_i - z_j)^2.$$

Therefore, the necessary and sufficient condition for a reduction in polarization due to adding weight δ to edge (i,j) is:

$$-\partial_{w_{ij}} P(L) > (z_i - z_j)^2 \frac{\mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}}{\delta^{-1} + \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}},$$

which amounts to a steep enough first derivative.

Lemma 2.3 also allows us to study when polarization *increases* after adding weight to edge (i,j). In particular, if $\frac{\widetilde{\mathbf{z}}^T (I+L)^{-1} \mathbf{v}_{ij}}{\widetilde{\mathbf{z}}^T \mathbf{v}_{ij}} = 0$, then $P(\mathbf{z}^+) \geq P(\mathbf{z})$. Notice that if $\mathbf{v}_{ij}$ is not an eigenvector of L, then the addition of (i,j) can increase polarization when the mean-centered innate opinions $\widetilde{\mathbf{s}}$ lie on the $(n-1)$-dimensional subspace orthogonal to $(I+L)^{-2}\mathbf{v}_{ij}$. This condition is sufficient – but not necessary – the example in Figure 2.1 illustrates this point. Therefore, the planner cannot add edge weight arbitrarily and expect polarization to be reduced – the innate opinions can determine the sign of the effect.

However, there are special cases in which polarization is always reduced, such as the following.

Corollary 2.5. *If $\mathcal{G}$, i, and j satisfy $N(i) = N(j)$, then polarization is always reduced by adding weight δ to the edge (i,j), and the difference is*

$$P(\mathbf{z}) - P(\mathbf{z}^+) = (z_i - z_j)^2 \frac{2\delta(1 + \delta + d_i - w_{ij})}{(1 + 2\delta + d_i - w_{ij})^2}.$$

This result follows from proving that $L\mathbf{v}_{ij} = (d_i - w_{ij})\mathbf{v}_{ij}$ under the assumptions; see Appendix 2.A for full details.

Corollary 2.5 is somewhat counter-intuitive – if we strengthen connections between individuals who share the same set of neighbors, we may expect to form an 'echo chamber'. However, the opinion dynamics show that the addition of weight to such an edge (i,j) will *only* affect the expressed opinions of vertices i and j. While this edge fails to have any global effect, it does indeed bring the opinions of its incident vertices closer together – hence reducing polarization. The limitation of these effects to only its incident vertices suggests that in practice, the return on polarization may be small.

Lemma 2.3 is also used for arriving at the main result of this Section.

Theorem 2.6. *Let* $\mathbf{z}$, $\mathbf{z}^+$, δ, *and* $\mathbf{v}_{ij}$ *be as before. Then,*

$$P(\mathbf{z}) - P(\mathbf{z}^+) \leq \frac{1+\lambda_n(L)}{1+2\delta+\lambda_n(L)} \left(-\delta \partial_{w_{ij}} P(L)\right).$$

Furthermore, if there exists $\epsilon > 0$ *for which*

$$\frac{\widetilde{\mathbf{z}}^T (I+L)^{-1} \mathbf{v}_{ij}}{\widetilde{\mathbf{z}}^T \mathbf{v}_{ij}} \geq \epsilon + \frac{\delta}{2\delta + (1+\lambda_2(L))^2},$$

then we also have

$$P(\mathbf{z}) - P(\mathbf{z}^+) \geq \frac{2\delta\epsilon(z_i - z_j)^2}{1+2\delta}.$$

Theorem 2.6 directly motivates two heuristics for the planner. First, we see that the largest possible reduction in polarization is proportional to the first order effect $-\delta \partial_{w_{ij}} P(L)$. Therefore, it is natural for the planner to iteratively add maximal edge weight along the direction of steepest descent – a heuristic well-known as a *coordinate descent*. Additionally, for fixed ϵ, the lower bound grows with the expressed disagreement. Therefore, edges with large $(z_i - z_j)^2$ are also good candidates for the planner to add weight to; we name this strategy *disagreement-seeking*.

The upper bound in Theorem 2.6 implies that there is a diminishing return in adding more weight to a single edge, as $\frac{1+\lambda_n(L)}{1+2\delta+\lambda_n(L)} < 1$. In particular, this shows that although $P(\mathbf{z})$ is not globally convex, it is indeed convex along the direction of w_{ij}.

2.3.3 Adversarial Opinions

In some cases, the planner may not reliably use the innate or expressed opinions. For instance, they may be difficult (even impossible) to measure, or vertices may be susceptible to takeovers; see Gionis et al. (2013); Matakos et al. (2017, 2020); Gaitonde et al. (2020); Rahaman and Hosein (2021); Chen and Rácz (2022) for examples of the latter. Moreover, individuals' opinions may be multidimensional – capturing many distinct issues (e.g., firearm regulation, universal basic income, healthcare, etc.), all of which are shaped by the network's structure. Such cases may require the planner to take a robust approach: they seek to design a network structure that minimizes polarization for *any* possible vector of innate opinions.[5] Formally, they aim to solve the following:

[5]There is one other possible justification for this formulation – a robust (or minimax) optimization problem arises when the decision-maker is ambiguity averse, as is shown axiomatically by Gilboa and Schmeidler (1989a).

$$\begin{aligned} \min_{\mathcal{G}'} \max_{\mathbf{s}\in\mathbb{R}^n:\|\mathbf{s}\|_2^2\leq R} & \widetilde{\mathbf{s}}^T \left(I + L'\right)^{-2} \widetilde{\mathbf{s}} \\ & \text{s.t. } ||W - W'||_0 \leq 2k. \end{aligned} \tag{2.5}$$

Polarization in the resulting graph $\mathcal{G}'$ will be robust to the choice of innate opinions, and this optimization problem yields different graph structures than problem (2.2). As before, the factor of two in the constraint captures all graphs being undirected.

This optimization problem can be interpreted as a game – an adversary selects $\mathbf{s}$ from the n-dimensional sphere of radius R, and the planner evaluates polarization on this choice of $\mathbf{s}$. A similar problem appears in Gaitonde et al. (2020), who studies a 'network defender' that decreases vertices' susceptibility to the adversary. In contrast, we consider defending the network through modification of its structure. However, we note that both our defender and theirs face the same adversary. This choice allows us to directly compare the effectiveness of these two defensive strategies. Although such an adversary may not be realistic, we believe this setting has numerous other justifications.

Note that the innate opinions now lie in the n-dimensional sphere, as opposed to the hypercube. This formulation allows us to relate the adversary's problem to spectral properties of the resultant graph $\mathcal{G}'$. In fact, the planner's problem (2.5) is equivalent to maximization of λ_2, the spectral gap of the Laplacian.

Proposition 2.7. *The optimal solution $\mathcal{G}'$ to* (2.5) *is the same as that of*

$$\begin{aligned} \max_{\mathcal{G}'} \ & \lambda_2(L') \\ \text{s.t. } & ||W - W'||_0 \leq 2k, \end{aligned} \tag{2.6}$$

If the optimal solution to (2.6) *is L^*, then the optimal value of* (2.5) *is $\frac{R}{(1+\lambda_2(L^*))^2}$.*

For two graphs $\mathcal{G}$ and $\mathcal{G}'$, if $W \leq W'$ elementwise, then $L \preccurlyeq L'$, and therefore $\lambda_2(L) \leq \lambda_2(L')$. Therefore, the planner must only add edge weight to $\mathcal{G}$, as reducing weights cannot increase the spectral gap.[6] The spectral gap of the Laplacian is intimately tied to the synchronizability of various types of dynamical systems and the mixing time of Markov chains (Donetti et al., 2006) and hence several studies seek to maximize it (Watanabe and Masuda, 2010; Wang and Van Mieghem, 2010).

[6]We remark that this monotonicity of the spectral gap in the edge set does not hold for the normalized Laplacian $\mathcal{L}$, see for instance Eldan et al. (2017).

In this adversarial setting, where perfect synchronization is impossible, the spectral gap controls the best achievable consensus.

The proof of Proposition 2.7 follows from solving the inner maximization problem, for which the optimal solution is the eigenvector of L' corresponding to the second-smallest eigenvalue. This eigenvector is called the *Fiedler* vector of $\mathcal{G}'$, and describes a partition of vertices that approximates the normalized sparsest cut of $\mathcal{G}$ (Chung, 1997).

For a graph with Laplacian L, Proposition 2.7 indicates that the *worst-case* polarization is equal to $P(L) = \frac{R}{(1+\lambda_2(L))^2}$. The adversary achieves this by choosing $\widetilde{\mathbf{s}}$ along the span of the Fiedler vector. The planner's effectiveness in problem (2.5) is controlled by $P(L) - P(L')$, the difference in *worst-case* polarization.

As in the previous setting, we approach this problem by iteratively choosing edges to saturate – starting from the initial graph until no further budget remains. Therefore, the principal results address how increasing an edge's weight affects the spectral gap and thereby polarization. This is quantified in Theorem 2.8, which relates changes in the spectral gap to elementwise differences in the Fiedler vector.

Theorem 2.8. *Let $\mathcal{G}$ be an undirected graph, and (i, j) be an edge with non-maximal weight, that is, $w_{ij} < \bar{w}$. Let also $\mathbf{v}$ be the Fiedler vector of $\mathcal{G}$ of unit magnitude with corresponding eigenvalue $\lambda_2(L)$. Recall that $\lambda_3(L)$ is the third smallest eigenvalue of L, and define $\beta = \lambda_3(L) - \lambda_2(L)$.*

For some $\delta \in (0, \bar{w} - w_{ij}]$, let $\mathcal{G}^+$ be constructed by adding weight δ to edge (i, j). If $\alpha = |v_i - v_j|$, then we have that

$$\max\left\{1 - \frac{2\delta}{\beta}, 0\right\} \delta\alpha^2 \leq \lambda_2(L^+) - \lambda_2(L) \leq \delta\alpha^2.$$

The proof follows from adapting the result of Maas (1987). The bounds are tightest when β is largest, equivalently when $\lambda_2(L)$ is the sole small eigenvalue of L.

This result motivates a simple heuristic for maximizing the spectral gap, which appears in Wang and Van Mieghem (2010). The planner can iteratively compute the Fiedler vector and add weight to non-saturated edges whose incident vertices have large absolute difference in $\mathbf{v}$.

Corollary 2.9 quantifies the effects on polarization induced by the perturbation in Theorem 2.8.

Corollary 2.9. *Let $P(L) = \frac{R}{(1+\lambda_2(L))^2}$ be the worst-case polarization on a graph with Laplacian L. In the setting of Theorem 2.8, we have*

$$\frac{2R\delta}{(1+2\delta+\lambda_2(L))^3}\max\left\{1-\frac{2\delta}{\beta},0\right\}\alpha^2 \leq P(L)-P(L^+) \leq \frac{4R\left(\delta\vee\delta^2\right)}{(1+\lambda_2(L))^3}\alpha^2$$

In contrast to the setting with full information, the worst-case polarization $P(L)$ cannot increase when the planner increases an edge's weight. Recall that this follows from the monotonicity of the spectral gap in W. However, it is possible that the resulting graph $\mathcal{G}^+$ has greater polarization for *some* particular innate opinions. The settings in (2.2) and (2.5) are distinct, and therefore the quantities compared before and after edge-weight addition are fundamentally different.

2.4 Empirical Simulations

If we solved problems (2.2) or (2.6) naively, it would be necessary to test all $\sum_{i=1}^{k}\binom{\binom{n}{2}}{i}$ possibilities. Given that computing polarization (or the spectral gap) requires $O(n^3)$ time, we obtain a crude upper bound of $O(kn^{2k+3})$. Note that for fixed k, this rate is polynomial in n – albeit still not scalable. However, in subsequent experiments we choose k to grow linearly with n, which results in superexponential runtime. It is therefore extremely impractical to compute the optimal solution, and we resort to theoretically motivated heuristics.

In Sections 2.3.2 and 2.3.3, we briefly discussed three heuristics for solving the planner's problem in a greedy, iterative fashion. Our theoretical results studied how polarization is reduced by *addition* of weight to a single edge. Therefore, all of the following heuristics are based on increasing weights of edges in $E^C := \{(i,j) : w_{ij} < \bar{w}\}$. These are presented below – detailing the edge to be saturated (i.e. setting edge weight to $\bar{w}$) at every step and briefly discussing the time complexity of each iteration. We will compare these approaches with two baselines: adding random non-edges, and distributing a single unit of edge weight evenly among all non-edges.

- **Mixing K_n (M):** Add $\frac{k}{|E^C|}$ units of weight to each non-edge. This is equivalent to creating a graph whose adjacency matrix is a convex combination of the original (simple) A and that of complete graph $A_{K_n} = \mathbf{1}\mathbf{1}^T - I$. Namely, we study $\frac{k}{|E^C|}A_{K_n} + \left(1-\frac{k}{|E^C|}\right)A$. This approach can be implemented in constant time for any k.

- **Random:** Fully saturate an edge from E^C chosen uniformly at random; this has runtime of $O(\log(n))$.

- **Disagreement Seeking (DS):** $\underset{(i,j)\in E^C}{\operatorname{argmax}}\ (\bar{w}-w_{ij})(z_i-z_j)^2$.

Computing the expressed opinions requires $O(n^3)$ time, and it takes $O(|E^C|)$ time to check all candidate nonedges.

- **Coordinate Descent (CD):** $\underset{(i,j)\in E^C}{\text{argmax}} \; -(\bar{w} - w_{ij})\partial_{w_{ij}} P(\mathbf{z})$

  Requires $O(n^3)$ runtime for computing a matrix inverse and multiplication, and $O(|E^C|)$ to find the optimal edge.[7]

- **Fiedler Difference (FD):** $\underset{(i,j)\in E^C}{\text{argmax}} \; (\bar{w} - w_{ij})|v_i - v_j|$, where $\lambda_2 \mathbf{v} = L\mathbf{v}$

  Takes $O(n^3)$ time to compute the eigendecomposition of L, and $O(|E^C|)$ to find the argmax.

Notice two effects at play: the maximal weight that can be added ($\bar{w} - w_{ij}$), and some measure of effectiveness per unit weight (disagreement, partial derivative, or absolute difference in Fiedler vector). Naturally, each heuristic attempts to maximize the two's product.

In addition, note that the three non-baseline heuristics have total runtime of $O(k(n^3 + |E^C|))$. The random and MK baselines have shortest runtimes of only $O(k \log(n))$ and $O(1)$, respectively. However, computing polarization at each step (for purposes of comparison) comes with an additional cost of $O(kn^3)$. We believe that the random heuristic is useful for two reasons. First, it captures a totally naive recommendation system, which does not curate a user's content exposure based on their opinions. Second, in two of the random graph models we study – Erdős-Rényi and stochastic block – the result of the random heuristic is another graph from the same model, but with slightly higher edge density. Therefore, this heuristic allows us to study how much *additional* polarization is reduced by adding edges in an informed, targeted, manner. The MK baseline is useful for comparing the relative effectiveness of *targeted* vs *global* interventions on the network. It is equivalent to assuming that users on a platform are randomly and uniformly exposed to others' content – see Appendix 2.B for greater discussion and analysis.

We now study the performance of these heuristics on six unweighted graphs. First we look at three real-world networks – sourced from Twitter, Reddit, and political blogs – and then three synthetic networks with different characteristics: the Erdős-Rényi, stochastic block, and preferential attachment models. Table 2.1 provides basic information about the graphs studied. In what follows, the planner's budget is given by $k = \lfloor \frac{n}{2} \rfloor$, such that on average each vertex receives one new edge. We plot the value of polarization with the planner's budget, along with the reference point $P(\mathbf{z}_{K_n})$, which represents the global minimum of polarization.

[7] Naively, one might think we need $O(n^3|E^C|)$ time to find the optimum, as we perform a matrix multiplication to compute the gradient of every candidate edge. However, the matrix multiplication is extremely sparse, and can be reduced to operating on four entries of a fixed, pre-computed matrix.

Network	Vertices n	Edges m
Twitter	548	3638
Reddit	556	8969
Blogs	1222	16717
Erdős-Rényi	1000	9990
SBM	1000	13726
PA	1000	4883

Table 2.1: Initial networks for evaluation of polarization-reducing heuristics.

Table 2.2 shows three quantities: expressed polarization, spectral gap, and assortativity of innate opinions. Expressed polarization is the principal concern of this study, and through Proposition 2.7 is closely related to the spectral gap. Assortativity is introduced by Newman (2003), and captures the degree of homophily in a network – which has been shown to control the speed of consensus-forming (Golub and Jackson, 2012). In particular, assortativity lies in $[-1, 1]$, and measures the correlation of an attribute across edges. In these experiments, assortativity is evaluated for the innate opinions.

Consistently, the random baseline decreases polarization the least, which is closely followed by the MK baseline. Both the DS and CD heuristics outperform the Fiedler vector-based strategy. This is expected, as the FD heuristic is blind to the innate opinions, and uses strictly less information. However, we observe that DS and CD tend to result in negative values of homophily, while the FD heuristic does not share this tendency. As an interesting implication, it does not appear that a reduction in polarization requires negative values of homophily. Namely, it may not be necessary to directly connect the most polarized individuals in a society to reduce its level of polarization.

In the figures that follow, vertices are colored according to their innate opinions. Graphs are plotted using the python module `networkx` (Hagberg et al., 2008). Vertices are placed in two-dimensional space using force-directed algorithms, in which vertices repel each other and edges behave like springs in tension. Therefore, the vertex layout reflects their relative attraction. The same random seed for initial node placement is used for every graph type studied. All code and data used to produce these results is publicly available here.

2.4.1 Real-World Networks

The Twitter and Reddit datasets used in this section were first collected by De et al. (2014), and used by both Chen and Rácz (2022) and Musco et al. (2018) in recent work. An additional dataset comprised of political blogs was collected by Adamic and Glance (2005) and used in Matakos et al. (2017, 2020).

Quantity	Heuristic	Real-World Networks			Synthetic Networks		
		Twitter	Reddit	Blogs	Erdős-Rényi	SBM	PA
Expressed Polarization	*Initial*	0.166	0.0053	36.6	0.242	3.53	1.71
	MK	0.086	0.0032	17.9	0.214	2.52	1.24
	Random	0.101	0.0035	22.1	0.219	2.58	1.35
	DS	0.022	**0.0006**	**8.2**	0.143	**1.77**	0.62
	CD	**0.020**	**0.0006**	**8.2**	**0.142**	**1.77**	**0.61**
	FD	0.075	0.0013	15.1	0.201	1.81	1.23
Spectral Gap	*Initial*	0.44	0.96	0.17	7.4	4.6	2.8
	MK	1.46	2.02	1.19	8.4	5.6	3.9
	Random	0.69	0.98	0.30	8.2	5.51	3.2
	DS	0.79	0.97	1.39	7.4	6.7	3.1
	CD	0.80	2.82	1.26	7.4	6.8	3.2
	FD	**2.05**	**9.17**	**2.33**	**12.0**	**6.9**	**4.0**
Assortativity of Innate Opinions	*Initial*	0.023	-0.007	0.811	-0.016	0.687	0.025
	MK	–	–	–	–	–	–
	Random	0.018	-0.005	0.779	-0.015	0.661	0.029
	DS	-0.143	-0.142	0.747	-0.114	0.606	-0.138
	CD	-0.090	-0.093	0.747	-0.102	0.618	-0.114
	FD	0.029	-0.007	0.780	-0.013	0.635	0.026

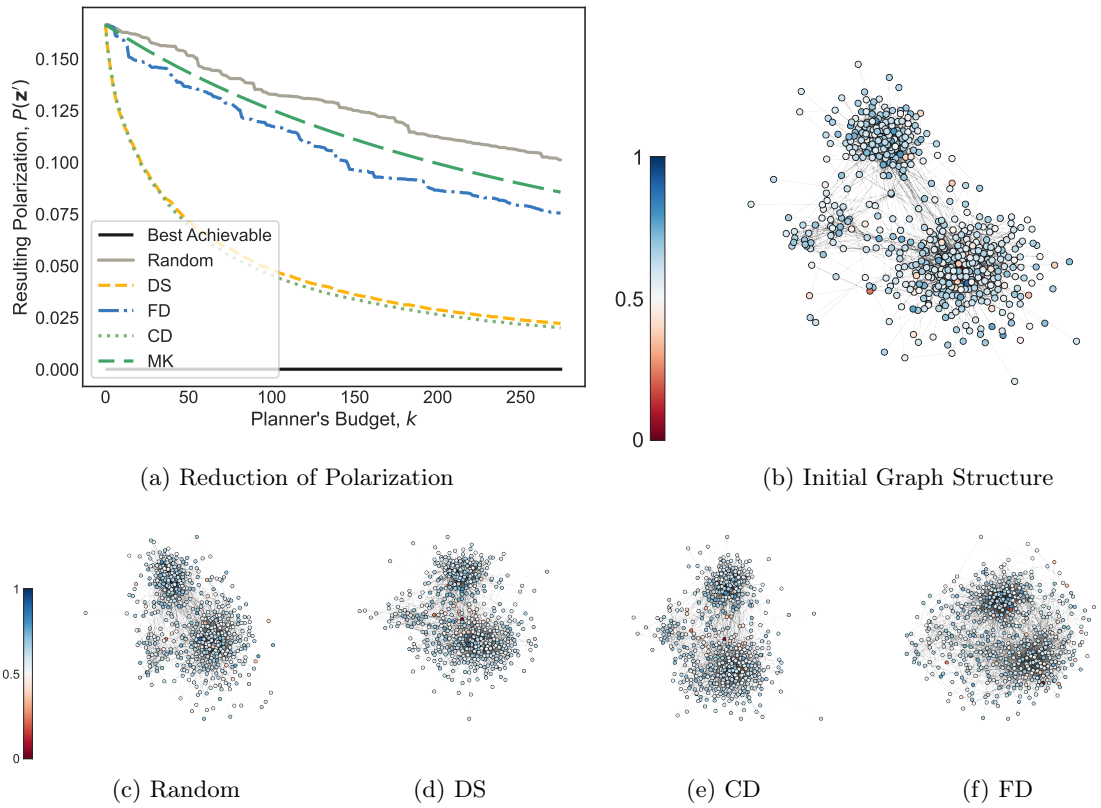
Table 2.2: Values for expressed polarization, spectral gap, and innate assortativity computed before and after the planner applies each heuristic to six networks. With the exception of the MK heuristic (see Appendix 2.B for more details on this approach), the planner adds $k = \lfloor \frac{n}{2} \rfloor$ edges – an average of one new edge per vertex. The best-performing heuristics are highlighted in bold. Assortativity depends only on the sparsity structure of the network, and is therefore not reported for the MK heuristic. Appendix 2.C contains additional figures showing changes in the spectral gap and assortativity with the planner's budget.

Twitter: This network reflects individuals who tweeted about a Delhi assembly debate in 2013. The network is shown in Fig. 2.2b, and mainly consists of two communities.

Fig. 2.2a shows the reduction in polarization achieved by the planner when applying each of the heuristics. Notably, all heuristics outperform our simple baselines. For the two best-performing heuristics, the first 50 edges modified reduce polarization by about a factor of two, and the subsequent 50 achieve a similar fractional reduction. This highlights both the substantial effect that the planner can have with minimal modifications to the graph, along with the diminishing returns of their budget.

The networks resulting from the planner's heuristics are shown in Figs. 2.2c-f. There are notable reductions in the strength of community structures. While less effective in reducing polarization, the Fiedler vector-based heuristic (FD) appears to smooth out communities the most.

Reddit: This network was generated by following Reddit users who posted in a politics forum. Three isolated vertices are removed in preprocessing. Fig. 2.3b shows that the initial network



(a) Reduction of Polarization

(b) Initial Graph Structure

(c) Random

(d) DS

(e) CD

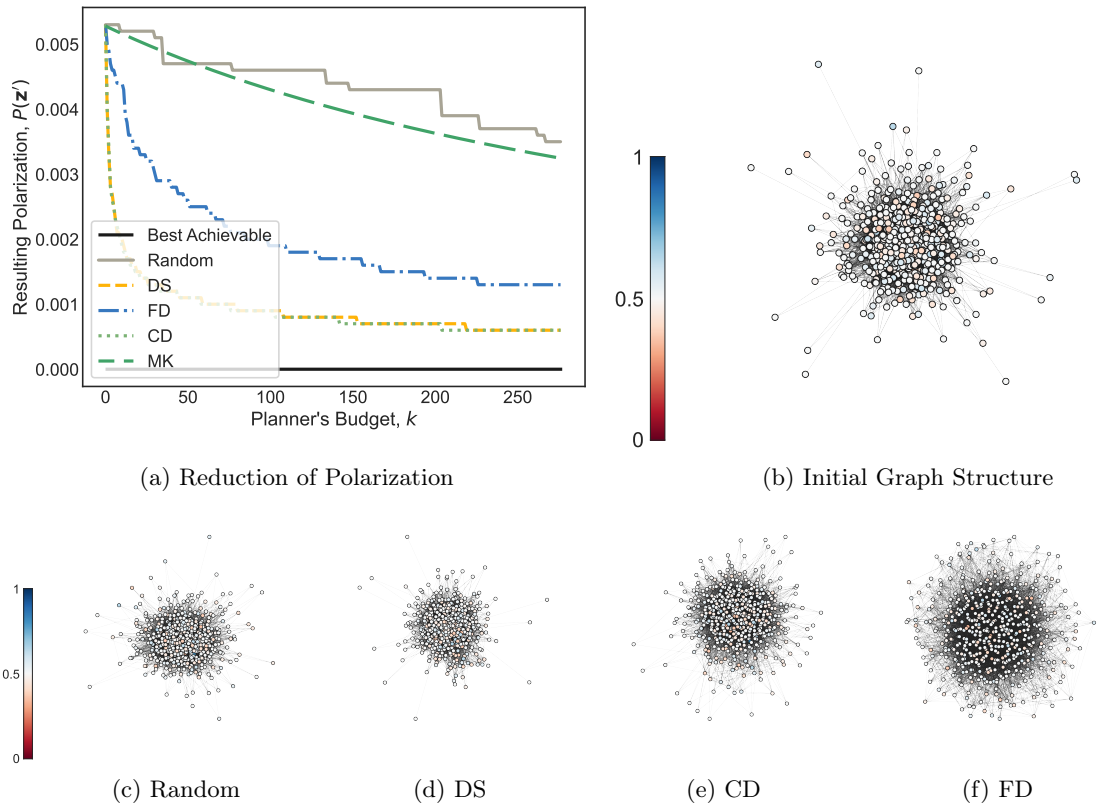
(f) FD

Figure 2.2: Evaluation of the planner's heuristics on the Twitter network. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)-(f) visualize the network after the planner has exhausted their budget according to each heuristic. Vertices are colored according to their innate opinions.

appears to be tightly clustered, and Table 2.2 indicates that it exhibits an extremely small level of polarization.

For any non-baseline heuristic, the full budget reduces polarization by almost a factor of four. For the best-performing heuristics, this reduction is by nearly an order of magnitude. We observe greatly diminishing returns, with the most significant reduction achieved with the first few edges modified. Moreover, the best-performing heuristics come close to achieving the globally optimal solution after fully exhausting the budget.

Only minor changes are observed in the resulting graph structures. Figs. 2.3c, 2.3d, and 2.3e look almost identical to the initial network. In contrast, the graph in Fig. 2.3f does not have as dense a core, and appears to be more evenly connected. Since maximizing the spectral gap results in the graph behaving similarly to an expander, which (informally) is equally well-connected across all cuts, this is to be expected.



(a) Reduction of Polarization

(b) Initial Graph Structure

(c) Random

(d) DS

(e) CD

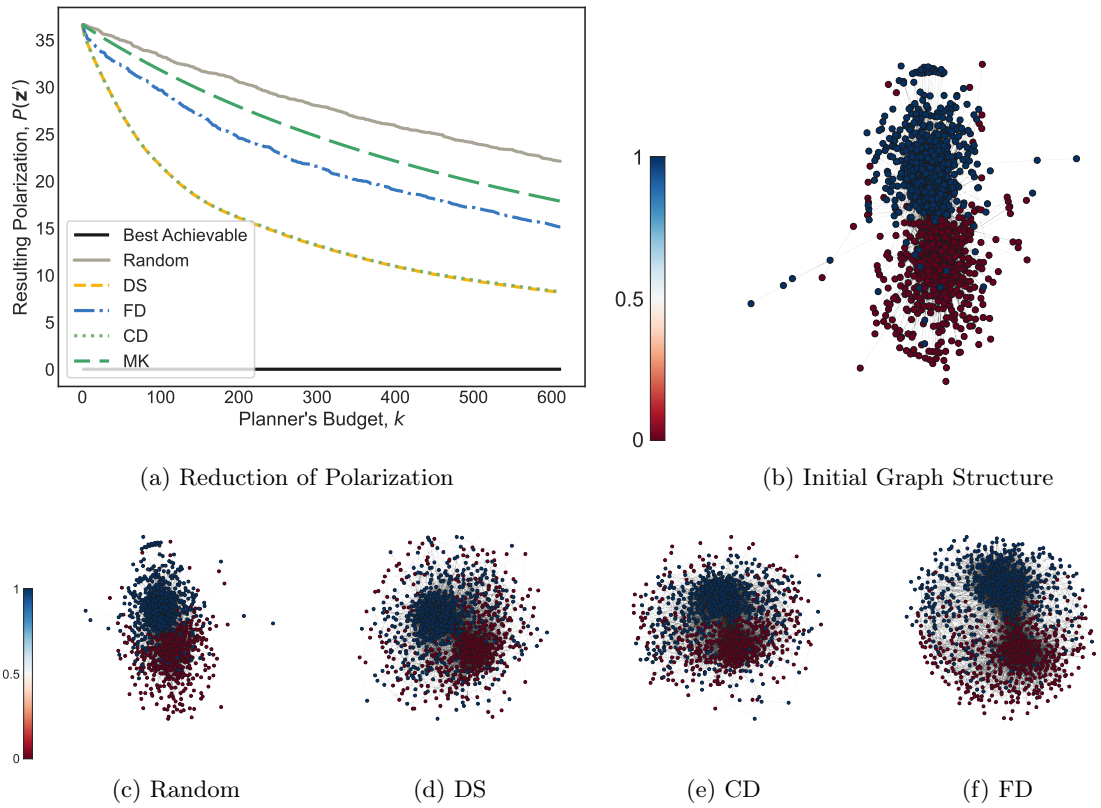
(f) FD

Figure 2.3: Evaluation of the planner's heuristics on the Reddit network. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)-(f) visualize the network after the planner has exhausted their budget according to each heuristic. Vertices are colored according to their innate opinions.

Blogs: This network was collected by aggregating online directories of political blogs around the 2004 US elections. Note that vertices in this network represent blogs – not individuals as in the previous datasets. Each blog was identified as either 'conservative' or 'liberal', which we encode by innate opinions of 0 or 1, respectively. Observe in Table 2.2 that this network exhibits extremely large values of polarization and homophily, and a small spectral gap.

We find consistent reductions in polarization with all heuristics – including the baselines. This network is unique in that the community structure largely mirrors the innate opinions. That is, the mean-centered innate opinions vector is highly collinear with the Fiedler vector. Hence, both the DS and FD heuristics will choose to modify edges between the two communities. Furthermore, a large fraction of the non-edges span the two communities – a randomly chosen edge is therefore likely to bridge the two.

Fig. 2.4c shares with Fig. 2.4b a tightly-knit core, with a few vertices at the extremities. In contrast, Figures 2.4d-f depict networks that are more uniformly connected. As before, we find



(a) Reduction of Polarization

(b) Initial Graph Structure

(c) Random

(d) DS

(e) CD

(f) FD

Figure 2.4: Evaluation of the planner's heuristics on the political blogs network. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)-(f) visualize the network after the planner has exhausted their budget according to each heuristic. Vertices are colored according to their innate opinions.

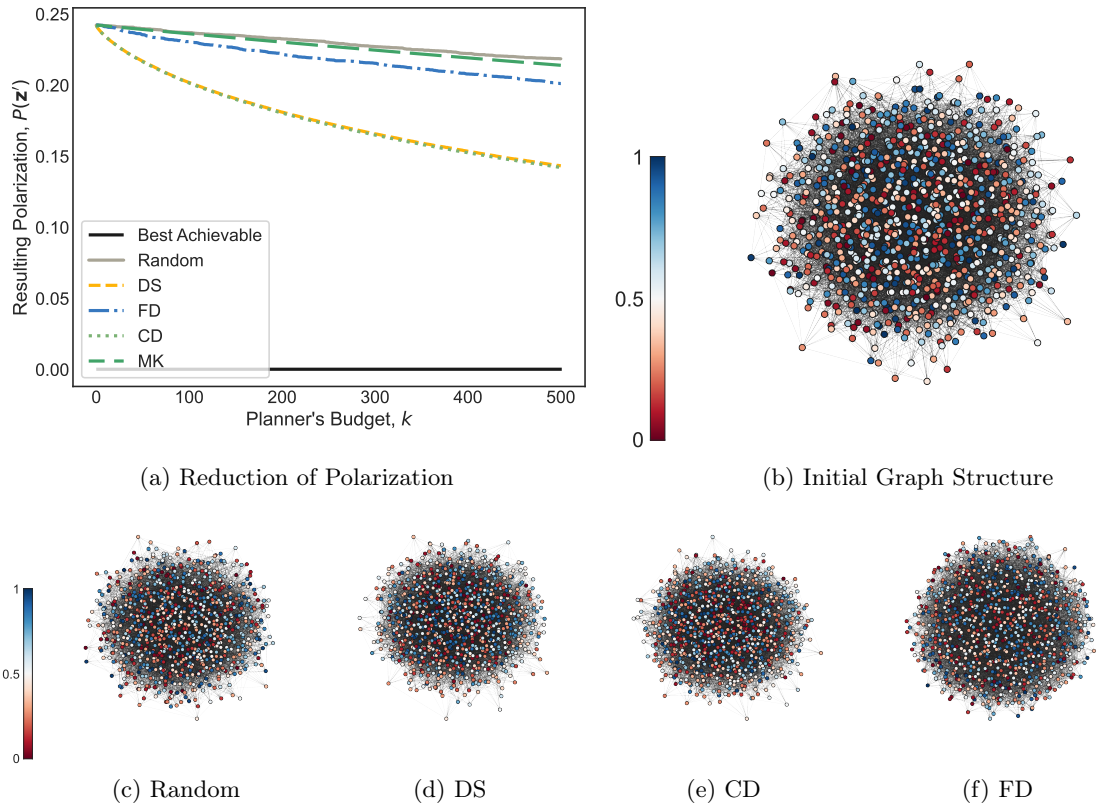
feature this to be most observable with use of the FD heuristic.

2.4.2 Synthetic Datasets

These heuristics are also applied to three canonical models of random graphs. In the first two models, the number of edges grows quadratically in the number of vertices (for fixed parameters). However, in our final model, the number of edges is always linear in the number of vertices. Therefore, one may expect that the impact of the planner's $O(n)$ edges is greatest in the sparser model – but we will see that this is not the case.

Erdős-Rényi: A graph from this model connects each pair of vertices independently with a fixed probability $p \in [0, 1]$. We take $n = 1000$ and $p = 0.02$, although the results are qualitatively similar for different values. The innate opinions are independent uniform random variables in $[0, 1]$.

This model produces homogeneous, well-connected networks, which are good spectral approxi-



(a) Reduction of Polarization

(b) Initial Graph Structure

(c) Random

(d) DS

(e) CD

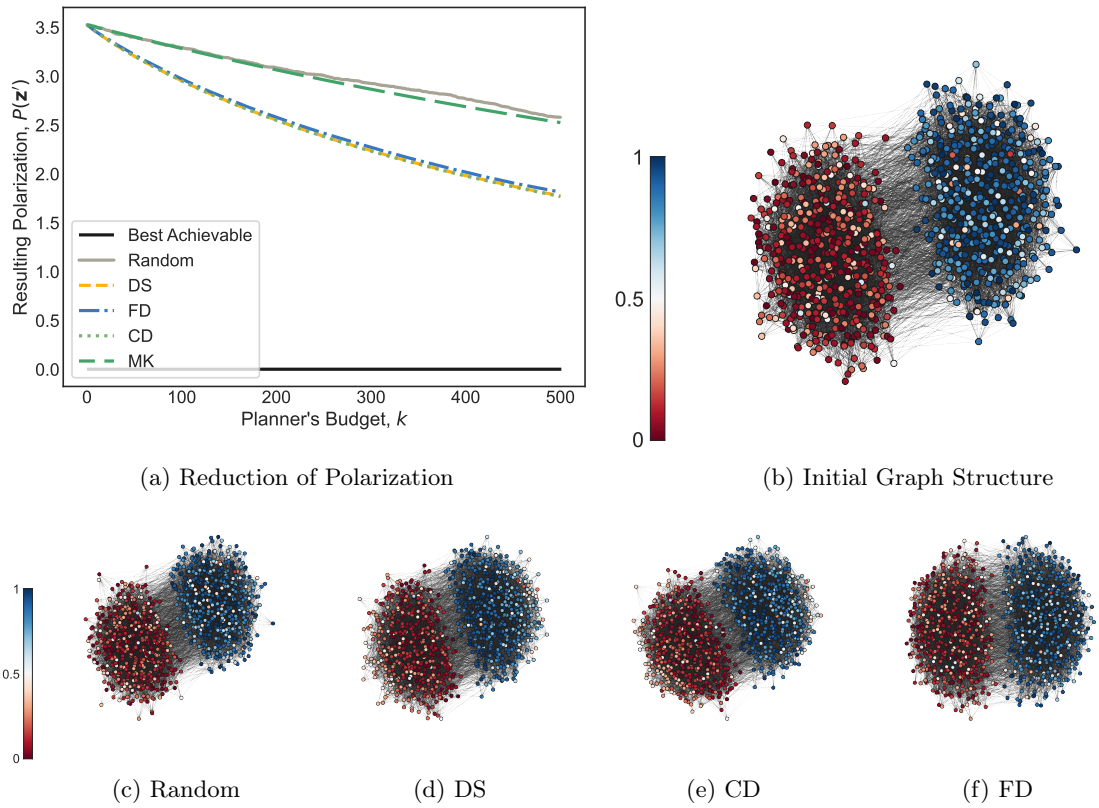
(f) FD

Figure 2.5: Evaluation of the planner's heuristics on the Erdős-Rényi graph. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)-(f) visualize the network after the planner has exhausted their budget according to each heuristic. Vertices are colored according to their innate opinions.

mations of the complete graph K_n (Hoffman et al., 2021). This can be seen through the large initial spectral gap in Table 2.2. Therefore, according to Proposition 2.1 it is natural to expect polarization to be small. Nonetheless, all heuristics fail to significantly reduce polarization. A comparison with the random baseline is particularly interesting in this model, as it results in another Erdős-Rényi graph, but with a slightly larger value of p.

Few changes can be seen among Figs. 2.5c-f. However, there are few vertices with extreme opinions in the fringes of Fig. 2.5d. Instead, these vertices tend to be concentrated in the center of the graph. This aligns with the most negative assortativity seen in Table 2.2. Notably, this feature is not visible in Figures 2.5f or 2.5c.

Stochastic Block Model: A graph drawn from a stochastic block model can replicate community structures, and is shown in Fig. 2.6b. This random graph on $n = 1000$ vertices with two equal-sized communities is generated by mirroring the methodology in Chen and Rácz (2022). Specifically, the probability of inter-community edges is given by $q = 0.005$, and the probability of intra-community



(a) Reduction of Polarization

(b) Initial Graph Structure

(c) Random

(d) DS

(e) CD

(f) FD

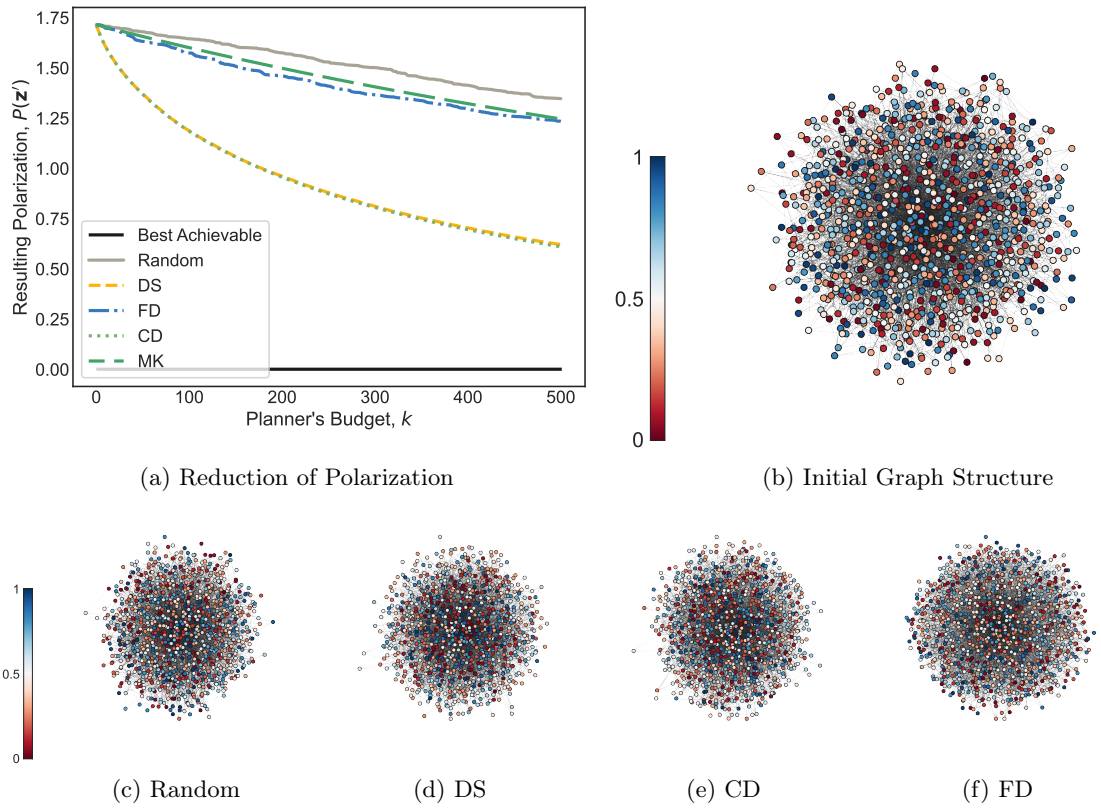
Figure 2.6: Evaluation of the planner's heuristics on the stochastic block model graph. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)-(f) visualize the network after the planner has exhausted their budget according to each heuristic. Vertices are colored according to their innate opinions.

edges is $p = 0.05$. Since $p > q$, we expect to see strong communities. The innate opinions of vertices in each community are drawn independently from either Beta$(1, 5)$ or Beta$(5, 1)$, such that the distribution of opinions mirrors the graph's community structure.

In Fig. 2.6a, a nearly identical reduction in polarization can be seen for all non-random heuristics. This occurs because the mean-centered innate opinions are highly collinear with the Fiedler vector, which partitions the graph into its two communities. Therefore, both the DS and FD strategies will add edges between the two communities. Similarly to the Erdős-Rényi setting, the random baseline yields another stochastic block graph, but with slightly larger parameters p and q.

Qualitatively, all three heuristics can be seen to bring the two communities closer together. However, in Fig. 2.6d and 2.6e, the vertices with extreme opinions are brought closer to the center. As before, this is not observed in Fig. 2.6f.

Preferential Attachment Model: This model generates graphs with power-law degree distribution, often known as scale-free or Barabási-Albert networks (Barabási and Albert, 1999). Again,



(a) Reduction of Polarization

(b) Initial Graph Structure

(c) Random

(d) DS

(e) CD

(f) FD

Figure 2.7: Evaluation of the planner's heuristics on the preferential attachment graph. Panel (a) shows the reduction achieved as the planner gradually adds edges. Panel (b) shows the initial network, while (c)-(f) visualize the network after the planner has exhausted their budget according to each heuristic. Vertices are colored according to their innate opinions.

we follow a similar procedure to Chen and Rácz (2022), with $n = 1000$ vertices added sequentially. Each incoming vertex connects to at most $m = 5$ vertices, where the likelihood of connecting to a particular vertex is proportional to its degree.

This graph tends to exhibit a small, highly interconnected core, and many vertices with low degree. This structure is not conducive to low polarization, as we see in Fig. 2.7a. The best-performing heuristics manage to reduce polarization by just over a factor of two, whereas the others see only negligible fractional reductions. Notably, the FD heuristic only slightly outperforms the baseline. These observations are a result of the Friedkin-Johnsen model – higher-degree nodes experience the smallest marginal effects of increased edge weight. Since the preferential attachment model yields a heavy-tailed degree distribution, a larger fraction of nodes are resilient to the planner's modifications. These nodes will also exert large amounts of influence on their neighbors due to their high degree. We therefore believe that the structural properties of the preferential attachment graph dampen the planner's effectiveness.

Qualitatively, Figs. 2.7d-f appear similar to the original network in Fig. 2.7b. We do not see strong changes in the structure, which aligns with the minor differences in homophily and spectral gap in Table 2.2.

2.5 Discussion and Conclusion

In this chapter, we analyze the relationship between structures of social and information networks and opinion polarization.

First, we establish a relationship between the ratio of *expressed* to *innate* polarization. This ratio is controlled by structural properties of the graph, such as the degree profile and isoperimetric number (i.e. Cheeger constant). In particular, the worst-case polarization depends directly on the spectral gap of the Laplacian. Consequentially, we show that the complete graph achieves the global minimum for polarization. This result aligns with one's intuition – bottlenecks in the graph are liabilities to a consensus.

Next, we present two variations of the planner's problem – one in which the innate opinions of the population are known, and another in which they are chosen adversarially. In the first, an expression is derived for the exact difference in polarization when weight is added to a single edge. We find that strengthening the connections between vertices with large *expressed* disagreement reduces polarization. However, it is seen as costly for individuals to interact with differently-minded others (Bindel et al., 2015). Therefore, reaching a consensus, while arguably beneficial for the population, may prove costly to individuals. We also present a second setting wherein the planner defends the network against adversarially-controlled innate opinions. Here, we prove that the planner aims to maximize the spectral gap. We then show the effectiveness of a strategy based on the Fiedler vector $\mathbf{v}$ – the eigenvector corresponding to the spectral gap. Intuitively, this vector partitions the graph based on the signs of its elements, and the planner should strengthen edges across the cut.

Finally, we evaluate the performance of four heuristics on several real-world and synthetic networks. We find that all strategies may smooth out community structures – often referred to as 'echo-chambers'. Furthermore, when there are no strong communities present, the Fiedler vector-based strategy is able to reduce polarization without simultaneously reducing homophily. With this approach, a reduction in polarization did not necessitate direct connections between opposite-minded individuals. However, this strategy performed significantly worse in several networks. We believe that the difference reflects how much of the polarization is driven by the particular values of opinions. For instance, all three heuristics perform similarly when the profile of opinions mirrors

the graph structure, and therefore both contribute similarly to the level of polarization. Specifically, all heuristics behave similarly when the mean-centered innate opinions $\widetilde{\mathbf{s}}$ are highly aligned with the Fiedler vector – this observation can be seen most easily in the blogs and stochastic block model networks.

There are several interesting directions for future theoretical work. First, this work has only derived bounds for single-edge modifications. It is an open problem to characterize the effects on polarization of more substantial perturbations to the graph structure. Furthermore, it may be possible to study the planner's effectiveness within various classes of random graph models. For instance, what fraction of non-edges in an Erdős-Rényi graph reduce polarization when added?

At first glance, the results in this chapter are severely limited by the model of opinion dynamics. Experimental research has shown that exposure to differing opinions may increase polarization (Bail et al., 2018). Motivated by these observations, many models incorporate non-attractive forces between opinions – see Cornacchia et al. (2020); Rahaman and Hosein (2021) for extensions of the FJ model, and Hegselmann et al. (2002); Hązła et al. (2019); Gaitonde et al. (2021) for geometrically-inspired approaches. Within the broader problem of understanding how social network structures relate to polarization, this work provides only a first step – the analytical tractability of the FJ model comes at the expense of expressibility. Nonetheless, we believe these results may be generalizable to a wider class of opinion dynamics models that exhibit attraction – which includes all of the above examples. For instance, one could modify a 'disagreement-seeking' heuristic to only consider non-saturated edges between individuals within each others' radius of attraction. The study of polarization-reducing strategies in these more complex models of opinion interaction is a rich and fruitful area for future work.

Several networks showed large reductions in polarization with a small number of edge modifications. However, in the Erdős-Rényi and preferential attachment networks, the heuristics presented in this work did not have as strong of a performance. Beyond our speculation, it remains to be understood what properties of these networks may limit the planner's effectiveness, or what minimal budget is necessary for a fixed fractional reduction in polarization. Moreover, it is not yet clear if this observation is a feature of the heuristics or the graph itself – are we closely approximating the true optimal solution?

In this study, we have shown that strengthening ties between disagreeing individuals is an effective strategy for reducing social polarization. Therefore, if polarization is instead increasing as society becomes increasingly connected, then both individuals and social media platforms may be failing to contribute to discourse between opposing perspectives.

Appendices

2.A Proofs

First, we specify notation. Let I denote the identity matrix, $\vec{\mathbf{1}}$ the all-ones vector, and $\mathbf{e}_i$ the i-th standard basis vector – all of appropriate dimension. Additionally, for $\mathbf{x} \in \mathbb{R}^n$, we write $\overline{x} := \frac{1}{n}\sum_{j=1}^{n} x_j$ to denote the mean of its entries and $\widetilde{\mathbf{x}} := \mathbf{x} - \overline{x}\vec{\mathbf{1}}$ for the mean-centered version of $\mathbf{x}$. For a square matrix $A \in \mathbb{R}^{n\times n}$, we write A_i for the i-th column of A. The eigenvalues of A in descending order are given by $\lambda_n(A) \geq \lambda_{n-1}(A) \geq \ldots \geq \lambda_1(A))$. We frequently use the notation $\lambda_{\max}(A) = \lambda_n(A)$ and $\lambda_{\min}(A) = \lambda_1(A)$ to denote the largest and smallest eigenvalue of A, respectively.

Given an initial graph $\mathcal{G}$ and any other graph $\mathcal{G}'$, define $T \equiv T(\mathcal{G}'; \mathcal{G}) \in \mathbb{R}^{n\times n}$ to be

$$T := (I + L)^{-1}(I + L'), \tag{2.7}$$

where L and L' denote the combinatorial Laplacians of $\mathcal{G}$ and $\mathcal{G}'$, respectively. The dependence of T on $\mathcal{G}$ and $\mathcal{G}'$ will be clear from context and hence omitted. The expressed opinions $\mathbf{z}'$ can be computed in terms of T and the original expressed opinions as follows:

$$\mathbf{z}' = T^{-1}\mathbf{z}.$$

This matrix is also useful in allowing us to express the new value of polarization in terms of the expressed opinions on the initial graph. After some algebra, we have that

$$P(\mathbf{z}') = \widetilde{\mathbf{z}}^T (T^{-1})^T T^{-1} \widetilde{\mathbf{z}},$$

where we used (2.7). The spectrum of T will be critical for theoretical results.

Recall the definition of the isoperimetric number (also known as the Cheeger constant) of a graph

from (2.1). The following simple Lemma is useful in many of the subsequent proofs.

Lemma 2.10. *Let $d_{\max}$ and $d_{\min}$ denote the maximum and minimum weighted degrees of $\mathcal{G}$. Additionally, let L be the combinatorial Laplacian of $\mathcal{G}$, and let $\lambda_n \geq \lambda_{n-1} \geq \ldots \geq \lambda_2 \geq \lambda_1 = 0$ denote its eigenvalues in decreasing order. Then, we have that*

$$\frac{1}{2} d_{\min} h_{\mathcal{G}}^2 \leq \lambda_2 \leq 2 d_{\max} h_{\mathcal{G}}, \tag{2.8}$$

and also

$$\lambda_n \leq (2 d_{\max}) \wedge (\bar{w} n). \tag{2.9}$$

Proof. For the normalized Laplacian $\mathcal{L}$, the well-known Cheeger inequality (see, e.g., Chung (1997)) gives that

$$\frac{h_{\mathcal{G}}^2}{2} \leq \lambda_2(\mathcal{L}) \leq 2 h_{\mathcal{G}}.$$

Notice that the eigenvalues of $L = D^{1/2} \mathcal{L} D^{1/2}$ are equal to those of $\mathcal{L}D$. Additionally, the ordered eigenvalues of D are simply the degrees of $\mathcal{G}$ in descending order. Since both $\mathcal{L}$ and D are positive semidefinite and Hermitian, we can apply a Weyl multiplicative inequality from Horn and Johnson (1994) to establish that

$$\lambda_{i+j-n}(\mathcal{L}D) \leq \lambda_i(\mathcal{L}) \lambda_j(D), \qquad \text{if} \quad i + j - n \geq 1 \tag{2.10}$$

and

$$\lambda_i(\mathcal{L}) \lambda_j(D) \leq \lambda_{i+j-1}(\mathcal{L}D), \qquad \text{if} \quad i + j - 1 \leq n. \tag{2.11}$$

Choosing $i = 2$ and $j = n$ in (2.10) gives that

$$\lambda_2(\mathcal{L}D) \leq \lambda_2(\mathcal{L}) d_{\max} \leq 2 h_{\mathcal{G}} d_{\max}.$$

With $i = 2$ and $j = 1$ in (2.11), we have that

$$\lambda_2(\mathcal{L}D) \geq \lambda_2(\mathcal{L}) d_{\min} \geq \frac{1}{2} h_{\mathcal{G}}^2 d_{\min}.$$

Combining the previous two displays gives (2.8).

The inequality (2.9) can be proved using the triangle inequality. The largest eigenvalue of $\mathcal{L}$ is equal to the operator norm of $D - A$. Since the norm of both D and A are upper bounded by $d_{\max}$, we conclude that $\lambda_n(L) \leq 2d_{\max}$.

Let $L_{K_n} = \bar{w}\left(nI - \vec{\mathbf{1}}\vec{\mathbf{1}}^T\right)$ denote the combinatorial Laplacian of the complete graph, where all edge weights are equal to $\bar{w}$. Since $L_{K_n} \succcurlyeq L$, for any L, we have that $\bar{w}n = \lambda_n(L_{K_n}) \geq \lambda_n(L)$. □

Proof of Proposition 2.1. We seek to write $P(\mathbf{z})$ in a way that $P(\mathbf{s})$ appears. Recall that $\mathbf{z} = (I + L)^{-1}\mathbf{s}$ and also $\widetilde{\mathbf{z}} = (I + L)^{-1}\widetilde{\mathbf{s}}$. Therefore

$$P(\mathbf{z}) = \widetilde{\mathbf{z}}^T\widetilde{\mathbf{z}} = \widetilde{\mathbf{s}}^T(I + L)^{-2}\widetilde{\mathbf{s}}. \tag{2.12}$$

Towards the lower bound in the claim, we may use an eigenvalue bound to obtain that

$$P(\mathbf{z}) \geq \lambda_{\min}((I + L)^{-2})\widetilde{\mathbf{s}}^T\widetilde{\mathbf{s}} = (1 + \lambda_{\max}(L))^{-2}P(\mathbf{s}).$$

From (2.9) we have that $\lambda_{\max}(L) \leq (2d_{\max}) \wedge \bar{w}n$; plugging this into the display above we obtain the claimed lower bound.

For the upper bound, first note that the eigenvector corresponding to the largest eigenvalue of $(I + L)^{-2}$ is $\vec{\mathbf{1}}$. Since $\widetilde{\mathbf{s}}$ is orthogonal to $\vec{\mathbf{1}}$, we have from (2.12) that

$$P(\mathbf{z}) \leq \lambda_{n-1}((I + L)^{-2})\widetilde{\mathbf{s}}^T\widetilde{\mathbf{s}} = (1 + \lambda_2(L))^{-2}P(\mathbf{s}).$$

From (2.8) we have that $\lambda_2(L) \geq (1/2)d_{\min}h_{\mathcal{G}}^2$; plugging this into the display above we obtain the desired upper bound. □

Proof of Corollary 2.2. Take any graph $\mathcal{G}$ and innate opinions $\mathbf{s}$. Proposition 2.1 implies that

$$P(\mathbf{z}_{\mathcal{G}}) \geq P(\mathbf{s})(1 + (2d_{\max}) \wedge (\bar{w}n))^{-2} \geq P(\mathbf{s})(1 + \bar{w}n)^{-2}.$$

Turning to the complete graph K_n, recall that the spectrum of its Laplacian has 0 as an eigenvalue with eigenvector $\vec{\mathbf{1}}$. It also has eigenvalue $\bar{w}n$ with multiplicity $n - 1$ and eigenspace containing all vectors orthogonal to $\vec{\mathbf{1}}$. Since $\widetilde{\mathbf{s}}^T\vec{\mathbf{1}} = 0$, we have $(I + L_{K_n})^{-1}\widetilde{\mathbf{s}} = (1 + \bar{w}n)^{-1}\widetilde{\mathbf{s}}$. Recalling the definition of polarization, we obtain $P(\mathbf{z}_{K_n}) = \left\|(I + L_{K_n})^{-1}\widetilde{\mathbf{s}}\right\|^2 = (1 + \bar{w}n)^{-2}\left\|\widetilde{\mathbf{s}}\right\|^2 = (1 + \bar{w}n)^{-2}P(\mathbf{s})$. Comparing with the display above, we see that K_n minimizes polarization over all graphs with maximal weight $\bar{w}$. □

Proof of Lemma 2.3. To obtain the claim, we expand $P(\mathbf{z}^+)$ in a way that $P(\mathbf{z})$ appears. First, note that $L^+ = L + \delta L_{ij}$ and $L_{ij} = \mathbf{v}_{ij}\mathbf{v}_{ij}^T$, and hence we have that $T = I + \delta(I+L)^{-1}\mathbf{v}_{ij}\mathbf{v}_{ij}^T$. The Sherman-Morrison formula thus gives that $T^{-1} = I - \frac{\delta}{1+\delta\mathbf{v}_{ij}^T(I+L)^{-1}\mathbf{v}_{ij}}(I+L)^{-1}\mathbf{v}_{ij}\mathbf{v}_{ij}^T$. Plugging this into the formula for polarization, we obtain that

$$\begin{aligned} P(\mathbf{z}^+) = \left\|T^{-1}\widetilde{\mathbf{z}}\right\|^2 &= \left\|\left(I - \frac{\delta}{1+\delta\mathbf{v}_{ij}^T(I+L)^{-1}\mathbf{v}_{ij}}(I+L)^{-1}\mathbf{v}_{ij}\mathbf{v}_{ij}^T\right)\widetilde{\mathbf{z}}\right\|^2 \\ &= \widetilde{\mathbf{z}}^T\widetilde{\mathbf{z}} - \frac{2\delta\widetilde{\mathbf{z}}^T(I+L)^{-1}\mathbf{v}_{ij}\mathbf{v}_{ij}^T\widetilde{\mathbf{z}}}{1+\delta\mathbf{v}_{ij}^T(I+L)^{-1}\mathbf{v}_{ij}} + (\widetilde{\mathbf{z}}^T\mathbf{v}_{ij})^2\frac{\delta^2\mathbf{v}_{ij}^T(I+L)^{-2}\mathbf{v}_{ij}}{\left(1+\delta\mathbf{v}_{ij}^T(I+L)^{-1}\mathbf{v}_{ij}\right)^2}. \end{aligned}$$

Noting that $P(\mathbf{z}) = \widetilde{\mathbf{z}}^T\widetilde{\mathbf{z}}$, and $D_{ij}(\mathbf{z}) = (\widetilde{\mathbf{z}}^T\mathbf{v}_{ij})^2$ leads to the desired expression after rearranging. □

Proof of Proposition 2.4. For simplicity of notation, let $A = I + L$. Then, for any $t > 0$, we have

$$\frac{P(\mathbf{z}_{L+tL_{ij}}) - P(\mathbf{z}_L)}{t} = \frac{\widetilde{\mathbf{s}}^T\left[\left(A + t\mathbf{v}_{ij}\mathbf{v}_{ij}^T\right)^{-2} - A^{-2}\right]\widetilde{\mathbf{s}}}{t}$$

Using the Sherman-Morrison formula, we can compute that

$$\begin{aligned} \left[\left(A + t\mathbf{v}_{ij}\mathbf{v}_{ij}^T\right)^{-1}\right]^2 &= \left[A^{-1} - \frac{tA^{-1}\mathbf{v}_{ij}\mathbf{v}_{ij}^TA^{-1}}{1+t\mathbf{v}_{ij}^TA^{-1}\mathbf{v}_{ij}}\right]^2 \\ &= A^{-2} - 2t\frac{A^{-2}\mathbf{v}_{ij}\mathbf{v}_{ij}^TA^{-1}}{1+t\mathbf{v}_{ij}^TA^{-1}\mathbf{v}_{ij}} + o(t) \end{aligned} \tag{2.13}$$

where $\frac{o(t)}{t} = o(1) \to_{t\to 0} 0$. Plugging (2.13) into (2.4) and taking the limit concludes. □

Proof of Corollary 2.5. Recall that $\mathbf{v}_{ij} = \mathbf{e}_i - \mathbf{e}_j$. Since $N(i) = N(j)$, a direct computation gives $L\mathbf{v}_{ij} = (d_i - w_{ij})\mathbf{v}_{ij}$. Consequently, we have $(I+L)^{-1}\mathbf{v}_{ij} = \frac{1}{1+d_i-w_{ij}}\mathbf{v}_{ij}$. Plugging this into Lemma 2.3 and simplifying yields the desired result. □

Proof of Theorem 2.6. The proof of this Theorem follows from bounding the terms in Lemma 2.3.

First, we show the upper bound. Notice that $\frac{\delta^2\mathbf{v}_{ij}^T(I+L)^{-2}\mathbf{v}_{ij}}{1+\delta\mathbf{v}_{ij}^T(I+L)^{-1}\mathbf{v}_{ij}} \geq 0$, so this term can be dropped. Through an eigenvalue bound we also find that

$$\frac{1}{1+\delta\mathbf{v}_{ij}^T(I+L)^{-1}\mathbf{v}_{ij}} \leq \frac{1}{1+2\delta\lambda_{\min}((I+L)^{-1})} = \frac{1+\lambda_n(L)}{1+2\delta+\lambda_n(L)}.$$

Plugging these two observations into (2.3) and rearranging to find $\partial_{w_{ij}}P(L)$ concludes.

For the lower bound, we have the following sequence of inequalities.

$$\frac{\delta \mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}}{1+\delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} \leq \frac{\delta \mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}}{1+\delta \mathbf{v}_{ij}^T (I+L)^{-2} \mathbf{v}_{ij}} \leq \frac{2\delta}{2\delta + (1+\lambda_2(L))^2}.$$

Therefore, by assumption and Lemma 2.3, we have that

$$P(\mathbf{z}) - P(\mathbf{z}^+) \leq \frac{\delta (z_i - z_j)^2}{1+\delta \mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij}} [2\epsilon] \leq \frac{2\delta\epsilon (z_i - z_j)^2}{1+2\delta},$$

where we used $\mathbf{v}_{ij}^T (I+L)^{-1} \mathbf{v}_{ij} \leq 2$. □

Proof of Proposition 2.7. This proof requires only that we solve explicitly the adversary's optimization problem.

By construction, $\widetilde{\mathbf{s}}$ is orthogonal to $\vec{\mathbf{1}}$. As a result, the optimal solution for the adversary is $\sqrt{R}\mathbf{v}_2$, where $\mathbf{v}_2$ is the eigenvector corresponding to the spectral gap of L', and their optimal value is:

$$\max_{\mathbf{s} \in \mathbb{R}^n : \|\mathbf{s}\|_2^2 \leq R} \widetilde{\mathbf{s}}^T \left(I + L'\right)^{-2} \widetilde{\mathbf{s}} = \frac{R}{\left(1+\lambda_2(L')\right)^2}.$$

To minimize this quantity, it follows that the planner maximizes the spectral gap of L'. □

Proof of Theorem 2.8. This proof uses a variation of a result by Maas (1987). The original result states that if a simple (unweighted, undirected) graph $\mathcal{G}_{\mathrm{s}}^+$ is constructed by adding a non-edge (i,j) to another simple graph $\mathcal{G}_{\mathrm{s}}$, we have

$$\min\left\{\lambda_2(L_{\mathrm{s}}) + \frac{\epsilon\alpha^2}{\epsilon + (2-\alpha^2)}, \lambda_3(L_{\mathrm{s}}) - \epsilon\right\} \leq \lambda_2(L_{\mathrm{s}}^+) \leq \min\{\lambda_2(L_{\mathrm{s}}) + \alpha^2, \lambda_3(L_{\mathrm{s}})\},$$

where $\alpha^2 = (v_i - v_j)^2$, and $\mathbf{v}$ is the eigenvector of L corresponding to $\lambda_2(L)$.

It is possible to adapt the original proof to consider the case where weight δ is added to edge (i,j). This result would yield:

$$\min\left\{\lambda_2(L) + \frac{\epsilon\delta\alpha^2}{\epsilon + \delta(2-\alpha^2)}, \lambda_3(L) - \epsilon\right\} \leq \lambda_2(L^+) \leq \min\{\lambda_2(L) + \delta\alpha^2, \lambda_3(L)\},$$

The tightest lower bound is achieved by choosing

$$\epsilon^* = \frac{\beta - 2\delta}{2} + \left(\left(\frac{\beta - 2\delta}{2}\right)^2 + \beta\delta(2-\alpha^2)\right)^{1/2},$$

where $\beta = \lambda_3(L) - \lambda_2(L)$, so that both terms in the minimum are equal. First, we note that $\epsilon^* \geq \beta - 2\delta$, with equality when $\alpha = 2$. Additionally, the first term in the minimum is increasing in ϵ, so therefore we have

$$\frac{\beta - 2\delta}{\beta}\delta\alpha^2 \leq \frac{(\beta - 2\delta)\delta\alpha^2}{\beta - 2\delta + \delta(2 - \alpha^2)} \leq \lambda_2(L^+) - \lambda_2(L) \leq \alpha^2,$$

as claimed since $\alpha^2 \geq 0$. Finally, note that $\lambda_2(L^+) \geq \lambda_2(L)$ as $L^+ - L = \delta L_{ij} \succcurlyeq 0$. □

Proof of Corollary 2.9. Recall that we defined $P(L) = \frac{R}{(1+\lambda_2(L))^2}$. We first prove the upper bound by using (2.8):

$$\begin{aligned}\frac{1}{(1+\lambda_2(L))^2} - \frac{1}{(1+\lambda_2(L^+))^2} &\leq \frac{1}{(1+\lambda_2(L))^2} - \frac{1}{(1+\lambda_2(L)+\delta\alpha^2)^2} \\ &\leq \frac{\delta^2\alpha^4 + 2(1+\lambda_2(L))\delta\alpha^2}{(1+\lambda_2(L))^2\,(1+\lambda_2(L)+\delta\alpha^2)^2}.\end{aligned}$$

Since $\delta\alpha^2 \geq 0$ and $\alpha^2 \leq 2 \leq 2(1+\lambda_2(L))$, we write $\alpha^4 \leq 2(1+\lambda_2(L))\alpha^2$, and arrive at

$$\frac{1}{(1+\lambda_2(L))^2} - \frac{1}{(1+\lambda_2(L^+))^2} \leq \frac{4\alpha^2\left(\delta \vee \delta^2\right)}{(1+\lambda_2(L))^3}$$

as claimed.

The lower bound follows similarly by (2.8). For simplicity of notation, let $c = \max\left\{1 - \frac{2\delta}{\beta}, 0\right\}$. Then,

$$\begin{aligned}\frac{1}{(1+\lambda_2(L))^2} - \frac{1}{(1+\lambda_2(L^+))^2} &\geq \frac{1}{(1+\lambda_2(L))^2} - \frac{1}{(1+\lambda_2(L)+c\delta\alpha^2)^2} \\ &\geq \frac{c^2\delta^2\alpha^4 + 2c\delta\alpha^2(1+\lambda_2(L))}{(1+\lambda_2(L))^2\,(1+\lambda_2(L)+c\delta\alpha^2)^2}.\end{aligned}$$

Observe that $c^2\delta^2\alpha^4 \geq 0$, so this term can be dropped. Furthermore, $c\delta\alpha^2 \leq 2\delta$, which gives us:

$$\frac{1}{(1+\lambda_2(L))^2} - \frac{1}{(1+\lambda_2(L^+))^2} \geq \frac{2c\delta\alpha^2}{(1+2\delta+\lambda_2(L))^3},$$

as desired. □

2.B Mixing K_n

In this section, we analyze the baseline heuristic used in Section 2.4 wherein we take a convex combination of the given network and the complete graph K_n. A notable feature of this approach is that it takes a *global* perspective towards reducing polarization – each non-edge in the network is equally increased by some small amount. We can imagine this being the result of a network administrator who occasionally exposes a user to a piece of content selected uniformly at random from the social network. This differs from the random heuristic in that it *slightly* increases an individual's exposure to *all* others, as opposed to significantly increasing their exposure to only a single other user.

Note that this procedure is directly motivated by the result of Corollary 2.2, wherein K_n is shown to be the global minimum for polarization. It is also valuable to note that this baseline is *not* feasible for problem 2.2, since it modifies a large number of edges – albeit by a small amount. Nonetheless, we can analyze its properties and compare its effectiveness.

Given an initial graph with Laplacian L, and for mixing coefficient $\eta \in [0, 1]$, this heuristic will yield a graph with the following Laplacian matrix:

$$L_\eta = \eta\bar{w}\left(nI - \mathbf{1}\mathbf{1}^T\right) + (1-\eta)L. \tag{2.14}$$

Recall that $\bar{w}\left(nI - \mathbf{1}\mathbf{1}^T\right) = L_{K_n}$, the Laplacian of a complete graph (with maximal edge weight $\bar{w}$).

For our empirical results, to establish a common scale we will choose the mixing coefficient η depending on k so that the graphs resulting from all heuristics have identical edge weight. Recall that we begin with a simple (i.e. unweighted) graph. Therefore, for some integer k, we have $\eta = \frac{k}{|E^C|}$, since weight η will be added to each of the $|E^C|$ non-edges.

This procedure is extremely convenient for analysis since both L_{K_n} and L share the same eigenbasis. We can show the following bounds on the resulting polarization, which match up to a factor of the condition number of L squared.

Proposition 2.11. *Let L denote the Laplacian of a graph, possibly weighted with maximal weight of $\bar{w}$, and $\mathbf{s}$ some innate opinions. For any $\eta \in [0, 1]$ and L_η as defined in* (2.14)*, let also $\mathbf{z}_\eta$ denote the expressed opinions corresponding to L_η. Then, we have:*

$$\left(\frac{1}{(1+\eta\bar{w}n)^2}\right) \vee \left(\frac{(1+\lambda_2(L))^2}{(1+\eta\bar{w}n+(1-\eta)\lambda_n(L))^2}\right) \leq \frac{P(\mathbf{z}_\eta)}{P(\mathbf{z})} \leq \frac{(1+\lambda_n(L))^2}{(1+\eta\bar{w}n+(1-\eta)\lambda_n(L))^2}.$$

Proof. This proof is relatively simple. First, let $L = U\Lambda U^T$ denote the decomposition of L, where U is orthonormal and Λ is diagonal.

We can write $L_{K_n} = \bar{w}\left(U(nI)U^T - \mathbf{1}\mathbf{1}^T\right)$, and therefore will have:

$$I + L_\eta = U\big[(1 + \eta\bar{w}n)I + (1-\eta)\Lambda\big]U^T - \eta\bar{w}\mathbf{1}\mathbf{1}^T.$$

Using the fact that $\mathbf{1}$ is an eigenvector of $U\big[(1 + \eta\bar{w}n)I + (1-\eta)\Lambda\big]U^T$, we can compute:

$$(I + L_\eta)^{-1} = U\big[(1 + \eta\bar{w}n)I + (1-\eta)\Lambda\big]^{-1}U^T - C\mathbf{1}\mathbf{1}^T,$$

for some constant C depending on η, n, and $\bar{w}$. Since $\widetilde{\mathbf{s}}$ is orthogonal to $\mathbf{1}$, we will have the following expression for the resulting polarization:

$$P(\mathbf{z}_\eta) = \widetilde{\mathbf{s}}^T U\big[(1 + \eta\bar{w}n)I + (1-\eta)\Lambda\big]^{-2}U^T\widetilde{\mathbf{s}}.$$

By definition, we will have $P(\mathbf{z}) = \widetilde{\mathbf{s}}^T U(I+\Lambda)^{-2}U^T\widetilde{\mathbf{s}}$, from which we obtain:

$$\frac{P(\mathbf{z}_\eta)}{P(\mathbf{z})} = \frac{\mathbf{y}^T(I+\Lambda)\big[(1 + \eta\bar{w}n)I + (1-\eta)\Lambda\big]^{-2}(I+\Lambda)\mathbf{y}}{\mathbf{y}^T\mathbf{y}}, \text{ with } \mathbf{y} = (I+\Lambda)^{-1}U^T\widetilde{\mathbf{s}}.$$

Straightforward bounds on this Rayleigh quotient give almost exactly the desired result:

$$\frac{1}{(1+\eta\bar{w}n)^2} \leq \frac{P(\mathbf{z}_\eta)}{P(\mathbf{z})} \leq \frac{(1+\lambda_n(L))^2}{(1+\eta\bar{w}n + (1-\eta)\lambda_n(L))^2}. \tag{2.15}$$

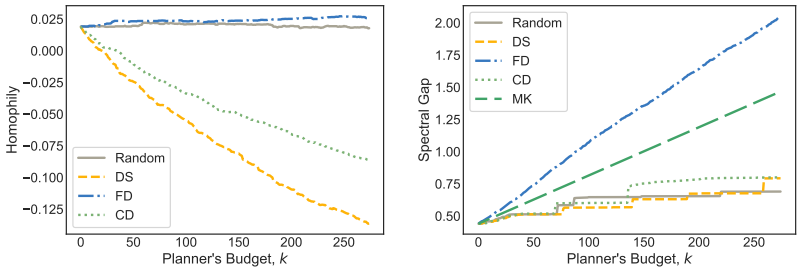
The second term appearing in the lower bound can be obtained by observing that $\lambda_n(L_\eta) = \eta\bar{w}n + (1-\eta)\lambda_n(L)$, and therefore:

$$P(\mathbf{z}_\eta) \geq \frac{P(\mathbf{s})}{(1+\lambda_n(L_\eta))^2} = \frac{P(\mathbf{s})}{(1+\eta\bar{w}n + (1-\eta)\lambda_n(L))^2}.$$

Using the inequality $P(\mathbf{s}) \geq P(\mathbf{z})(1+\lambda_2(L))^2$ and combining with the lower bound in (2.15) concludes. □

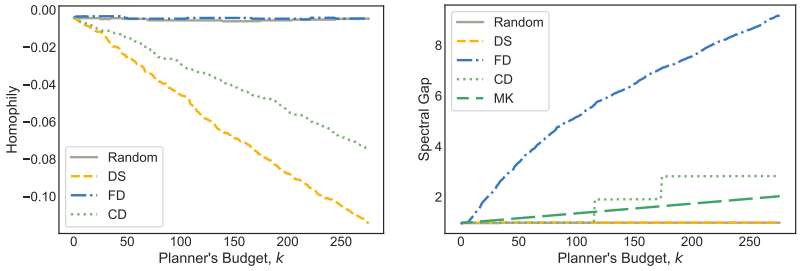
2.C Additional Figures

In this short section, we present Figures showing how homophily (i.e., assortativity of innate opinions) and the spectral gap are affected by the planner's modifications. These provide greater detail than the initial and final values found in Table 2.2.



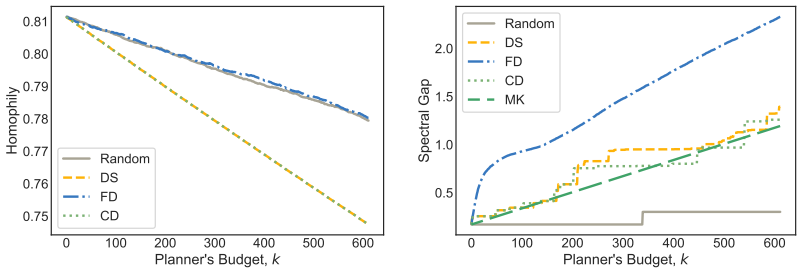
(a) Assortativity of innate opinions (b) Spectral gap

Figure 2.C.1: Impacts of the planner's budget on the Twitter network.



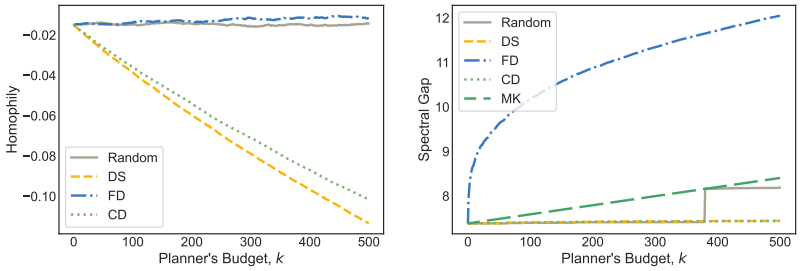
(a) Assortativity of innate opinions (b) Spectral gap

Figure 2.C.2: Impacts of the planner's budget on the Reddit network.



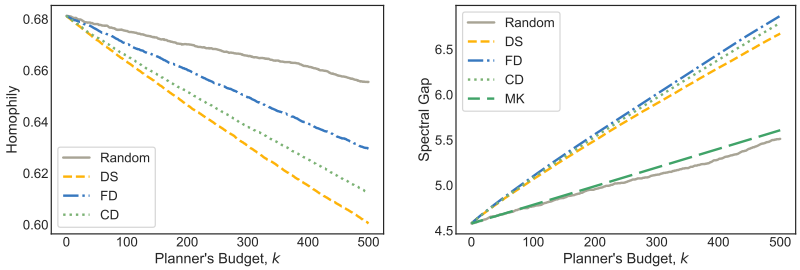
(a) Assortativity of innate opinions (b) Spectral gap

Figure 2.C.3: Impacts of the planner's budget on the political blogs network.



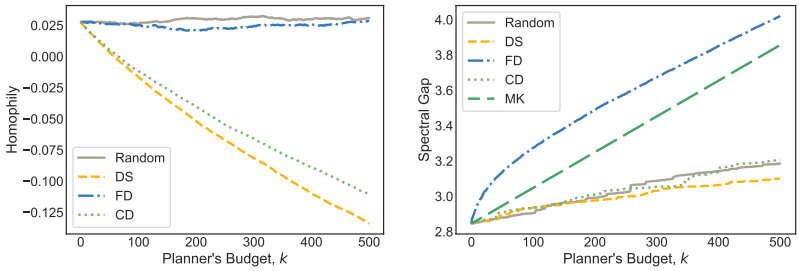
(a) Assortativity of innate opinions (b) Spectral gap

Figure 2.C.4: Impacts of the planner's budget on the Erdős-Rényi graph.



(a) Assortativity of innate opinions (b) Spectral gap

Figure 2.C.5: Impacts of the planner's budget on the stochastic block model graph.



(a) Assortativity of innate opinions (b) Spectral gap

Figure 2.C.6: Impacts of the planner's budget on the preferential attachment graph.

Chapter 3

Risks in Formation of Interbank Lending Networks

3.1 Introduction

Since the global financial crisis of 2008, systemic risk has become a topic of great interest to researchers, industry professionals, and policymakers alike. It is believed that interconnections between large financial institutions may have allowed distress to propagate throughout the financial system, and even beyond into other economic sectors. The crisis was extremely costly – even with bailouts of nearly $500 billion (provided by the US government), losses to the global economy totaled over $2 trillion (Lucas, 2019). This event renewed researchers' interest in understanding fragility of financial systems, and how policymakers can effectively intervene.

The phenomenon in which distress spreads through an entire system is dubbed 'systemic risk'. In financial applications, it is natural to study systemic risk through the perspective of networks and complex systems – the spread of distress is facilitated by the financial network that links institutions. Its edges represent, for instance, interbank loans (Allen and Gale, 2000) or overlapping portfolios (Cifuentes et al., 2005), and thereby highlight pathways along which distress can propagate. We may therefore expect that the observed patterns of financial contagion are related to this underlying network structure.

There is a substantial amount of research studying the dependence of systemic risk on characteristics of the financial network; for recent surveys see Jackson and Pernoud (2021) and Benoit et al. (2017). For example, a well-known contribution by Allen and Gale (2000) studies how sev-

eral stylized networks of direct interbank claims can yield different patterns of contagion – or even no contagion at all. The initial shock caused by idiosyncratic demand for liquidity can cause an overwhelmed banks' neighbors to suffer further liquidity shortages. A fully connected network is found to be optimal for sharing liquidity and therefore reducing the possibility of a systemic crisis. However, Gai and Kapadia (2010) identify a substantial tradeoff – the resilience of highly connected financial networks is accompanied by an increased intensity of systemic events. This feature, dubbed 'robust yet fragile', implies that although systemic crises are unlikely, they cause catastrophic and widespread damage.

A critical assumption in this branch of the literature is that the financial network is *exogenous*. Such papers are therefore restricted to analyzing the effect of a particular generative model for the network on systemic risk. Although these models serve as a useful Fbaseline, this assumption is unlikely to hold in practice. Instead, the network structure may be *endogenous*; each financial institution's connections reflect a set of optimal decisions. This perspective has become more prevalent in the literature, with relevant contributions by Bluhm et al. (2014), Acemoglu et al. (2015), and Farboodi (2021). Interestingly, it is possible for systemic risk to emerge hand-in-hand with each financial institution's selfish optimal behavior. However, these individually-optimal decisions need not maximize the collective well-being of the financial system. In such cases, as in this paper, financial institutions may be failing to internalize the negative effects of their decisions on the entire system. It is therefore of interest to analyze the severity of these negative externalities and how they might be remedied.

In this chapter, we study the formation of such a continuous-time interbank lending network when banks both face (and can insure against) idiosyncratic liquidity shocks. We formulate a system-wide optimization problem for both interbank exposures and insurance, in which the resulting network of interbank linkages indicates the channels for (and magnitude of) the propagation of financial distress.

The model proceeds as follows: consider a financial system comprised of a given number of banks. These banks may specialize in different activities; some collect a large number of deposits, whereas others specialize in revenue generation. This heterogeneity is modeled by unique, proprietary, investment opportunities (i.e. a portfolio of commercial loans) available to each bank. We assume that these opportunities are scalable, but are only accessible to their associated bank. The interbank lending mechanism in our model allows, for example, a deposit-collecting bank A to obtain the large returns of investment bank B's unique opportunities through a direct loan of capital from A to B – after which B invests this amount into their revenue-generating operations. In this setup, bank B is

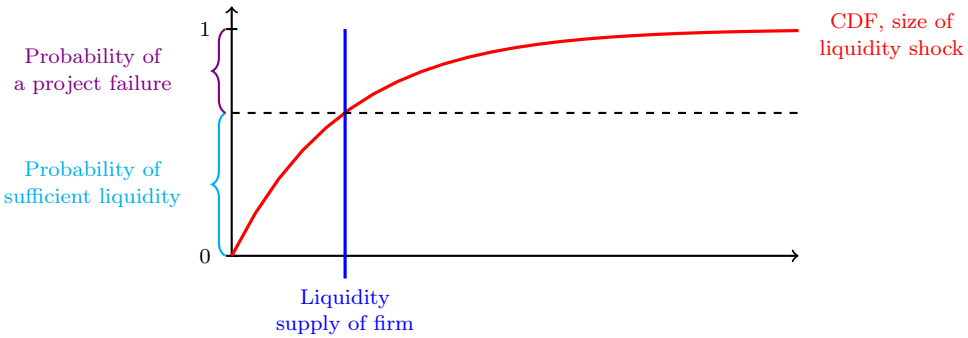


Figure 1: For a single bank, the relationship between the cumulative distribution function (CDF) of the size of a liquidity shock, their supply of cash, and the conditional probabilities of (in)sufficient liquidity.

effectively operating as an intermediary between bank A and B's own investment opportunities. We note that this construction is similar to the model of both Rochet and Tirole (1996) and Acemoglu et al. (2015), wherein banks invest in each other's 'projects' (henceforth, we will also use 'projects' to refer to these unique investment opportunities). In both these models and ours, the riskiness of these projects is tied to some decision taken by the associated bank.

Although these projects may accrue large rates of return, they are subject to a degree of risk. More precisely, a bank's project is periodically struck by liquidity shocks of random magnitude, and if the size of a shock exceeds the bank's current supply of liquidity (i.e. cash), then the project's value instantaneously drops (we refer to this event as a 'project's failure'). These shocks are assumed to represent, for example, additional liquidity required for the project to succeed, such as occurs in Rochet and Tirole (1996), Acemoglu et al. (2015), and more. If the required amount of cash is available, then the project continues smoothly. Conversely, a project's failure results in all its investors suffering losses proportional to their stake. Therefore, conditioned on the arrival of a liquidity shock, a bank's supply of cash determines their project's level of risk. Figure 1 illustrates the relationship between a bank's liquidity supply, the distribution of a liquidity shock's size, and the probabilities of each outcome.

In the model, bank A is assumed to have non-zero stake in their own project, and is therefore a co-investor of its creditors. Without this assumption, bank A would have no incentive to hold liquidity – they would be unaffected by their project's failure. In addition to holding cash, recall that bank A may lend their capital to any other bank B, which is invested into B's project. Banks in the system may also invest in a risk-free bond, or borrow at this rate from the central bank or external financiers. Finally, each bank is assumed to have some fixed amount of deposits, which fully specifies their balance sheet. An example is given in Figure 2, with descriptions of each item.

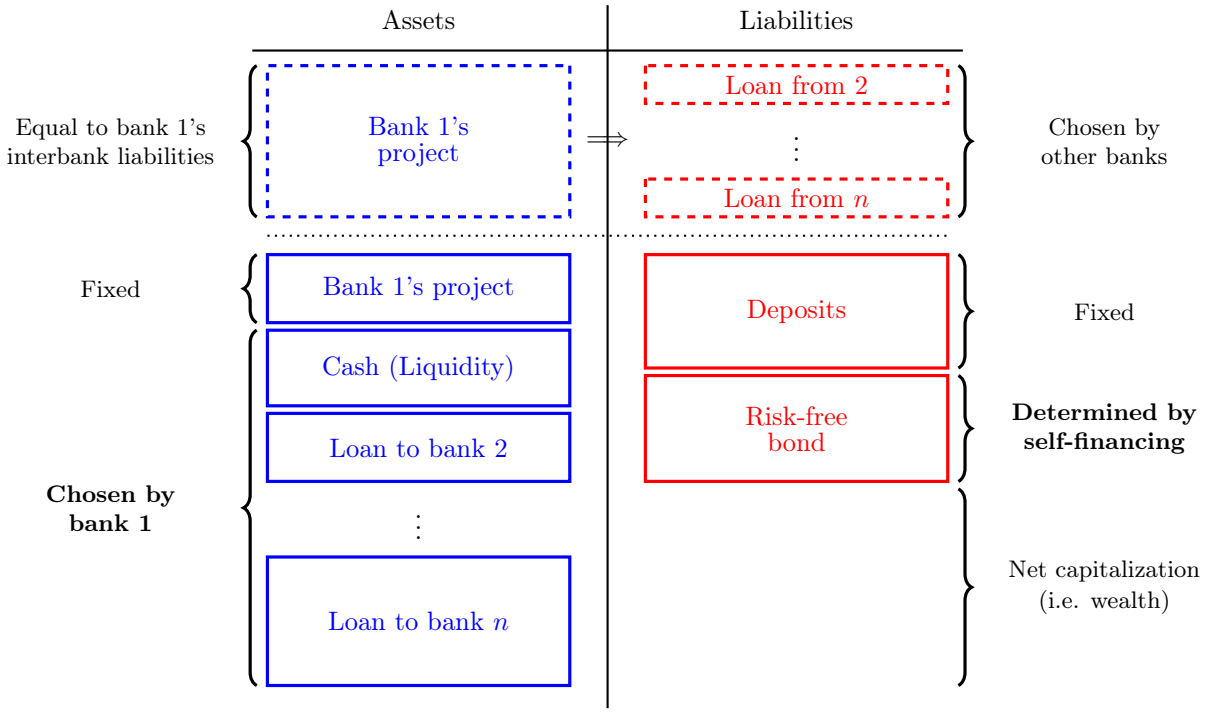


Figure 2: An example balance sheet for bank 1, who borrows at the risk-free rate to finance their investment portfolio. The decision variables for the bank are highlighted in bold.

A key focus of this work is that each bank endogenously chooses to allocate its capital between cash (i.e. supply of liquidity) and risky interbank loans. To that end, we will study the optimal capital allocations for two extreme settings of the financial system. First, consider the *decentralized* case – wherein each bank freely allocates their capital with pure self-interest. They seek only to maximize their utility of wealth at some terminal time. We note that this setting reflects a game-theoretic equilibrium. Second, we consider the *centralized* setting – where a single central planner makes the allocation decisions for all banks concurrently. The planner aims to maximize the sum of individual banks' utilities. In both of these cases, we derive the dynamic programming equations for the respective value functions, and explicitly compute the optimal allocations. Under stricter conditions, we can conclude uniqueness.

We observe a discrepancy between the optimal allocations computed in both settings; the central planner often chooses to hold a greater supply of liquidity. This occurs because our model captures a simple negative externality; when individualistically choosing their cash holdings, a bank determines the risk experienced by its creditors. An individual bank, operating in a selfish manner, fails to consider its creditors' losses when choosing their supply of cash. In contrast, the planner is cognizant of this systemic effect and acts accordingly by reducing the level of risk for banks with larger debt. Namely, the central planner achieves the welfare-maximizing (i.e. first-best) allocation for

the financial system. As a consequence of this discrepancy, a project's failure is more likely to occur with decentralized behavior than with a central planner. However, we also observe that the size of interbank loans is larger in the centralized setting, and hence each project's failure becomes more damaging to the system. This tension between the likelihood and severity of failures bears a resemblance to the 'robust yet fragile' observation made by Gai and Kapadia (2010). In particular, we find that this feature is associated with the socially optimal allocation of capital in the financial system.

We also study how the two optimal allocations differ as the financial system's size increases. Two natural points of comparison are: 1) the difference in, and 2) the ratio of, social welfare between both settings. The former comparison measures the nominal size of the inefficiency, and the latter its relative size (which has been dubbed the 'price of anarchy' by Papadimitriou (2001)). Perhaps counter-intuitively, we find that the price of anarchy remains bounded by a constant as the size of the system grows. Namely, the nominal size of the system's inefficiency grows at the same rate as the social welfare itself. These results are first derived theoretically, and also verified in simulations. Finally, we show that it is possible to alter banks' co-investment requirements to replicate the planner's optimal allocation.

There are several interesting consequences of our results. First, we find that the central planner's optimal allocation leads to low-frequency and high-intensity losses to the system. This may imply that the 'robust-yet-fragile' feature of financial networks is socially optimal. However, we see that the planner perfectly compensates for the larger-magnitude losses by ensuring they are less likely. As a result, the centralized allocation involves greater lending throughout the system. Additionally, in both settings we see that the (optimal) endogenous financial networks exhibit a strong 'core-periphery' structure, where only a subset of banks serve as borrowers to the rest of the system. Intuitively, we also show that systemically important banks must face the greatest losses if they are to replicate the planner's optimal allocation. This lends credence to the perverse incentives caused by 'too-big-to-fail' policies or other government bailouts.

This chapter is organized as follows. Section 3.1.1 reviews several relevant branches of literature. Section 3.2 introduces the model of interbank lending and the dynamics of each financial instrument. In the first part of our main results, Section 3.3 derives the optimal allocation in the decentralized (Section 3.3.1) and centralized (Section 3.3.2) settings. We compare these two optimal allocations in Section 3.4, including an asymptotic analysis of the price of anarchy. Finally, Section 3.5 concludes with a discussion of our results and directions for future work.

3.1.1 Related Literature

The foundational papers on continuous-time portfolio optimization are by Merton (1969, 1971). Merton studies the optimal portfolio allocation between risk-free and risky assets for a investor who maximizes their expected discounted utility of consumption. In these models, the returns of each risky asset are driven by correlated Brownian motions. Following from Merton's seminal papers, there is a wealth of literature on extensions of the original problem; see Rogers (2013) and references therein. The techniques we use in this work for deriving the optimal allocation will be similar to Merton's original work and its subsequent branch of literature. However, we will be studying a financial system in which *all* participants are simultaneously determining their optimal allocations of wealth – not only an individual. Moreover, to the best of our knowledge, the ability to control the jump intensity of a risky asset's returns has not been previously studied in the area of portfolio optimization.

There are, however, several papers that study an individual who incurs a cost to control the intensity of a jump process, such as Biais et al. (2010), Pagès and Possamaï (2014), Capponi and Frei (2015), Hernández Santibáñez et al. (2020), and Bensalem et al. (2020). These studies focus on Principal-Agent models and largely analyze the optimal contract and behavior. Moreover, they focus on the presence of moral hazard, where the Principal is unable to observe the Agent's efforts. Our mathematical approach for determining a bank's optimal supply of cash is similar to the models used in these papers. However, there are a few important differences. First, we study these optimizations performed simultaneously within a large system, and second, we focus on the inefficiencies that arise when individuals optimize greedily. Additionally, our setting assumes perfect information.

A strong motivation for this work follows from the systemic risk literature; much of the existing work assumes a given or exogenous network structure for the financial system. An early paper by Allen and Gale (2000) studies several stylized structures of interbank claims, and finds that the structure determines whether or not a local liquidity shock propagates throughout the system. More recent papers seek to answer similar questions with distinct models; for instance Gai and Kapadia (2010) and Gai et al. (2011) find that systemic liquidity crises can emerge in highly interconnected financial networks, albeit with low probability. Caccioli et al. (2014) present a model in which firms' overlapping portfolios can lead a single default to cause mark-to-market losses throughout the system – perhaps leading to additional defaults. In Elliott et al. (2014), firms directly own claims each others' assets and suffer sudden bankruptcy losses if their valuation falls below a threshold. Battiston et al. (2012) studies a continuous-time process representing financial robustness, and allows its evolution

to depend on a given financial network. Finally, several papers including Amini and Minca (2016); Detering et al. (2019, 2020) and Detering et al. (2021) seek to characterize the asymptotic behavior of contagion cascades in random inhomogeneous networks as the system's size grows. In their respective studies, these different mechanistic models are investigated both theoretically and in simulations. However, the explicit or implicit networks in these papers share one common feature – they are fixed or generated according to canonical random graph models. As previously highlighted, we believe this assumption may not be realistic; institutions in the financial system make optimal investment decisions, and the resulting network is endogenous – not random. In contrast to this branch of the literature, our model enables us to investigate how the organization and fundamental parameters of the financial system can lead to the emergence and scale of its inefficiencies.

We note that the high-level ideas in this chapter are similar to the literature on optimal network formation. For early work in this area, see Jackson and Wolinsky (1996) and Bala and Goyal (2000), where the authors present a process by which individuals choose to create edges with each other in a game-theoretic model. In these studies, individuals must balance a trade-off between the cost of forming an edge and the rewards associated with the edge. Our paper differs primarily from these studies through our emphasis on the financial features of the model, and edges are cost-less to form.

Most closely related to this work is the study of endogenous financial networks, including Zawadowski (2013); Bluhm et al. (2014); Acemoglu et al. (2015); Babus (2016) and Farboodi (2021). The work of Zawadowski (2013) shows that individual banks may fail to achieve the socially-optimal outcome by not buying insurance against their counterparties' default. While the author's model differs greatly from ours, we similarly see that individual banks' optimal behavior fails to internalize an externality on the system. A model by Babus (2016) presents an extension of Allen and Gale (2000). Her model allows banks to make optimal lending and borrowing decisions to redistribute liquidity throughout the system, and a highly-connected network is again found to be the most resilient to contagion. We share the idea of idiosyncratic liquidity shocks, but also study the planner's optimal allocation and compare it to the case where banks make selfish optimal decisions.

The three papers most similar to our own are Bluhm et al. (2014),Acemoglu et al. (2015) and Farboodi (2021). Our model is mechanistically different from the models in these studies – which are either static or consist of three distinct time periods. In contrast, we analyze the optimization problems in a dynamic continuous-time environment. First, Bluhm et al. (2014) construct a model of optimal interbank lending where banks face both liquidity and capital requirement constraints. In their model, both the interbank lending amounts and the market prices are determined endogenously. The authors show that contagion can occur (1) directly as a result of counterparty losses

in the event of a default, or (2) indirectly through the mark-to-market losses incurred by a bank's portfolio in the event of a fire sale. Despite the similarities to our paper, the authors largely focus on numerical and simulation results. In contrast, we seek to provide a theoretical characterization of the optimal solution wherever possible. Moreover, our model endogenizes the initial sources of disruption.

The contribution of Farboodi (2021) characterizes how banks optimally lend to each other within a financial system where there is a strong incentive to serve as intermediaries within the chain of lending. In her model, an interbank loan will also allow the lending bank to access the surplus generated by a risky investment of the borrowing bank. She shows that the resulting network can have a core-periphery structure, and that due to the benefit of intermediation, banks' private incentives can fail to achieve the socially optimal outcome. Although there are many similarities between this paper and ours, we do not focus on the incentive of intermediation, but instead on banks' optimal decisions to reduce the riskiness of their investments. Our results can, however, replicate the core-periphery feature in her paper – a small subset of banks with highly profitable investment opportunities form the financial network's core.

Finally, Acemoglu et al. (2015) endogenize both the decision of interbank lending and also the interbank interest rates. In a similar spirit to Rochet and Tirole (1996), banks exchange deposits to finance a project that yields high rate of return if run to conclusion, or low returns if liquidated prematurely. A bank faces external liabilities that may require them to liquidate these projects – thereby passing losses onto its creditors. The authors find that the optimal contracts do indeed consider the first-order network effects, wherein a risk-taking bank must pay large interest rates to its creditors. However, these do not account for the 'financial network externality', which can negatively affect banks that are not party to the contract. It follows that the resulting financial network may not be efficient (i.e. welfare-maximizing). While their model of interbank lending is similar to ours, the authors' analysis is largely focused on stylized networks in which equilibria are shown to exist. We will instead allow the sparsity structure of the financial network to be endogenously determined by the interbank lending opportunities.

3.2 Model

Consider a financial system consisting of n different banks. Let $(\Omega, \mathcal{E}, \mathbb{P})$ be a probability space, containing n independent Poisson processes $\tilde{N}_t^1, ..., \tilde{N}_t^n$, $t \geq 0$, each of which has corresponding intensity $\theta_1, ..., \theta_n > 0$. These counting processes will be used to indicate the arrival times of

liquidity shocks to each respective bank. Define $\mathcal{F}$ to be the filtration generated by the full set of jump processes. Hence, we obtain the filtered probability space $(\Omega, \mathcal{E}, \mathcal{F}, \mathbb{P})$.

The net capitalization (i.e. net value or wealth) of bank i is given by the non-negative stochastic process $\{X_t^i\}_{t\geq 0}$. We now aim to describe the dynamics of a bank's wealth.

3.2.1 Dynamics of Interbank Loans

First, the financial system contains a risk-free bond, which accumulates a constant, fixed rate of return r. Therefore its price, denoted S_t^0, evolves according to the ordinary differential equation $\frac{dS_t^0}{S_t^0} = rdt$. Banks may both borrow and invest at this risk-free rate, but in our model, we assume that an investment in the bond does not provide liquidity.

Each bank i has access to a unique set of external investments, e.g. a collection of commercial loans (henceforth referred to as a 'project'). These projects are available to another bank $j \neq i$ in the system through an interbank loan provided to i. In this manner, bank i serves as an intermediary between its project and the lending bank j. We will assume that there is no fee associated with this intermediation. Additionally, the capital invested in interbank loans is assumed to be illiquid. More precisely, neither the lending nor borrowing banks can use capital invested in a project to meet their liquidity needs. While these interbank claims indirectly accumulate large constant rates of return for investors, they will incur losses when the borrowing bank's project fails. If such a failure occurs, then the value of all capital invested in the project immediately drops. For instance, it is plausible that a bank's revenue operations intermittently require additional liquidity to cover a position or meet regulatory requirements. A failure to do so may lead to an inability to realize an investment's gains, or even directly cause losses.[1]

Let S_t^i denote the time-t value of a single unit of capital invested in bank i's project. Its dynamics are given by

$$\frac{dS_t^i}{S_t^i} = (\mu_i + r)\, dt - \phi_i\, dN_t^i, \quad i = 1, \ldots, n. \tag{3.1}$$

Observe that since $\mu_i > 0$, this interbank claim has rate of return larger than r. The jump increment dN_t^i is obtained by performing a thinning of the shock arrival process $\tilde{N}_t^i$, and is described in the next subsection. The increment takes on values in $\{0, 1\}$, and is non-zero if and only if bank i's

[1] There are several other distinct justifications for this feature of our model. For example, the bank may be investing in costly, continuous monitoring of its project, which reduces the risk of it suffering losses (e.g. default of its commercial loans). A different interpretation considers the effect of consumers' random liquidity preference. We may imagine that the shock represents depositors' demand to withdraw cash, and if the bank fails to meet this demand, they must liquidate the project at a loss to meet their more senior obligation.

project fails at time t. Finally, ϕ_i represents the magnitude of losses borne by investors in a project when it fails, i.e. $1 - \phi_i$ is the recovery rate.

Liquidity Shocks and Risk

All projects in the system may experience liquidity shocks; if sufficiently large, these shocks induce the project's failure. A key feature of this work is each bank's ability to control their project's susceptibility to failure – by holding a greater supply of liquidity, banks' projects are safer. In our model, this is represented through bank i's ability to influence the intensity of the jump increment dN_t^i that appears in (3.1).

A bank may hold a non-negative amount of their capital as cash, which has a constant price of 1. Although this capital effectively depreciates at the risk-free rate r (as it cannot be invested in the bond), it is the *only* source of liquidity within the system, and is the sole manner in which a bank can hedge against liquidity shocks. Namely, if a liquidity shock exceeds bank i's supply of cash, their project experiences a failure, and investors incur losses. The jump increment dN_t^i in (3.1) represents the arrival of shocks that overwhelm bank i's supply of cash. Its construction follows from a probabilistic model of liquidity shocks and a bank's supply of cash.

Recall that our filtered probability space contains n independent time-homogeneous Poisson processes $\tilde{N}_t^i$, with rates $\theta_i > 0$. At time t, if $\tilde{N}_t^i$ jumps, then bank i experiences a liquidity shock of size $X_t^i \zeta_t^i$, where the random variable ζ_t^i is $\mathcal{F}_t$-measurable. We assume that these shocks are proportional to a bank's wealth, and each ζ_t^i is independently and identically distributed according to the cumulative distribution function (CDF) $F_i(\cdot)$. The complementary CDF of ζ_t^i is defined as $\bar{F}_i(\cdot) = 1 - F_i(\cdot)$.

Let $c_t^i \geq 0$ denote the fraction of bank i's capital held in cash at time t. When the shock to bank i is larger than their supply of liquidity (i.e. $\zeta_t^i > c_t^i$), their project fails. When this occurs, all investors in the project suffer an instantaneous return of $-\phi_i$ on their investment amount. In particular, if $c_t^i = 0$, then any liquidity shock to bank i at time t, no matter how small, results in their project's failure.

The jump process N_t^i is constructed by independently flipping coins at every arrival time of $\tilde{N}_t^i$, with success probability given by $p_t^i = \bar{F}_i(c_t^i)$. Observe that $p_t^i = \mathbb{P}\left(\zeta_t^i > c_t^i \,\middle|\, \tilde{N}_t^i = 1\right)$. If and only if the flip is won, we let $dN_t^i = 1$. It follows that the instantaneous rate (at time t) of the Poisson process N_t^i is equal to $\theta_i \bar{F}_i\left(c_t^i\right)$.[2] The second component of the rate, $\bar{F}_i(c_t^i) = \mathbb{P}(\zeta_t^i > c_t^i)$, is the

[2]This result is a consequence of the thinning properties of Poisson processes. See, for instance, Theorem 1 in Lewis and Shedler (1979). If $\{c_t^i\}_{t\geq 0}$ is adapted (as we will require), then conditioned on time t, the previous jump process $\{N_s^i\}_{s\in[0,t]}$ has the desired rate function.

probability that bank i's project fails, conditional on the time-t arrival of a liquidity shock with CDF $F_i(\cdot)$. See Figure 1 for an illustration.

Finally, we will require a few technical conditions on F_i:

Assumption 3.1. *We assume that each F_i is absolutely continuous with respect to the Lebesgue measure. Its density is given by $f_i(\cdot) = F_i'(\cdot)$, which is assumed to be fully supported on $\mathbb{R}_+$, and monotonically decreasing (i.e. $f_i'(\cdot) < 0$).*

If $f_i(\cdot)$ had compact support, then it would be possible for a bank's project to be riskless with a large enough supply of liquidity. Since the return of this project would be greater than the risk-free rate, this would lead to all other banks to profit infinitely by borrowing at the risk-free rate and investing in the riskless project. While the problem may remain analytically tractable, this outcome is not of practical interest. Our assumption that the density is monotonically decreasing will be used to establish uniqueness of the optimal financial network.

3.2.2 Dynamics of Wealth

In this model, a bank may provide interbank loans to another; let $w_t^{ji} \geq 0$ denote the fraction of bank j's capital lent to bank $i \neq j$. The return experienced by this interbank claim is given by (3.1). Recall that c_t^i equals the fraction of bank i's wealth held as cash, which accumulates no return over time.

The final component influencing bank i's wealth is their degree of investment in their own project. We assume that each bank invests a fixed, given fraction of their current wealth. Unlike the interbank loans, we will assume that this quantity cannot be controlled.[3] This assumption has several possible interpretations. First, it may be the case that bank i is required by its creditors to be a co-investor in its project. We may also imagine that these projects are initialized by their respective banks, and simply scaled by any additional investments from the rest of the system. Therefore, the cost of initialization must be borne by the borrowing bank.

We will use $\frac{\eta_i}{\phi_i}$ to denote the fraction of i's wealth that is invested in their own project. This implies that bank i loses a constant fraction η_i of its total wealth whenever their project fails.[4] The parameter η_i captures the severity of a project's failure on the associated bank – in the extreme case of $\eta_i = 1$, a single failure will wipe out the bank i. Conversely, if $\eta_i = 0$, then bank i has no stake

[3] In principle, we could imagine allowing bank i to also control their exposure to their own project, while only being subjected to a minimum requirement. However, doing so introduces significant challenges in characterizing the optimal allocations.

[4] If, instead, this were a fixed amount and not fraction, then as a bank's wealth grows, their incentive to hold liquidity would become weaker. Such a setting is quite interesting in its own right, and may perhaps lead to a cyclic supply of liquidity – but it is not the focus of this chapter.

in their project and is unaffected by its failure. We will take $\eta_i \in (0,1)$, away from the two extreme cases. Therefore, the remaining $X_t^i(1 - c_t^i - \sum_{j\neq i} w_t^{ij} - \phi_i^{-1}\eta_i)$ units of wealth are invested in (or borrowed at) the risk-free rate.

Putting together the dynamics for each component of bank i's wealth, we see that X_t^i, follows

$$\frac{dX_t^i}{X_t^i} = \left(1 - c_t^i - \sum_{j\neq i} w_t^{ij} - \frac{\eta_i}{\phi_i}\right)\frac{dS_t^0}{S_t^0} + \sum_{j\neq i} w_t^{ij}\,\frac{dS_t^j}{S_t^j} - \frac{\eta_i}{\phi_i}\,\frac{dS_t^i}{S_t^i},\ i = 1, \cdots, n.$$

By using (3.1) and the dynamics of S_t^0, we obtain the following simplified expression:

$$\frac{dX_t^i}{X_t^i} = \left((1 - c_t^i)r + \sum_{j\neq i} w_t^{ij}\mu_j + \frac{\eta_i\mu_i}{\phi_i}\right)dt - \sum_{j\neq i} w_t^{ij}\phi_j\, dN_t^j - \eta_i\, dN_t^i,\ i = 1, \cdots, n. \tag{3.2}$$

A novel contribution of this work is the control c_t^i – while there is no return accumulated by this capital held as cash, it serves to reduce the likelihood that bank i's project fails, which would cause them to lose a fraction η_i of their wealth.

We say that $(c_\cdot^i, w_\cdot^{i\cdot}) \in \mathcal{A}_{s,t}^i$, the set of admissible controls for bank i between times s and t, if it is adapted to the filtration $\mathcal{F}$ and satisfies both $c_u^i \in \mathbb{R}_+$ and $w_u^{ij} \in [0, \phi_j^{-1})$ for all $u \in [s,t]$ and $j \neq i$. The upper bound on w_u^{ij} ensures that wealth will always remains positive.

All banks seek to maximize their own utility of wealth at a common terminal time $T < \infty$. As is relatively standard in the literature, a bank's utility function $U_i \in \mathcal{C}^\infty(\mathbb{R}_+)$ is assumed to have constant relative risk aversion:

$$U_i(x) = \begin{cases} \frac{x^{1-\gamma_i}}{1-\gamma_i} & \gamma_i > 0, \gamma_i \neq 1 \\ \log x & \gamma_i = 1. \end{cases} \tag{3.3}$$

3.3 Decentralized and Centralized Financial Networks

We consider two distinct organizations of the financial system. In the first, banks operate only in their self-interest – seeking to maximize their own expected terminal utility. We call this the *decentralized* setting, as there is no coordination between banks. Instead, each bank's optimal allocation reflects their best response to the others' decisions. On the other hand, the *centralized* setting in Section 3.3.2 will consider the perspective of a single central planner who determines all banks' allocations to maximize welfare – as measured by the sum of all banks' utilities.

Both allocations are important to consider. The decentralized optimum reflects a game-theoretic equilibrium of the financial system, where each bank chooses their controls optimally given all others' actions. Therefore, from the perspective of individual banks this is a stable allocation. In contrast, the centralized optimum reflects the maximum total utility that could exist in the financial system if banks coordinated. We will study the differences between these two optimal allocations, which reflect the severity of our model's externality, in Section 3.4. Finally, the optimal allocations yield a financial network of interest, which represents direct balance sheet exposures between banks.

3.3.1 Decentralized Network

Let us define the value function of bank i to be the supremum over all admissible controls of their expected utility at the terminal time:

$$V_i(t,x) = \sup_{(c^i,w^{i\cdot})\in\mathcal{A}^i_{t,T}} \mathbb{E}\left[U_i(X^i_T)\middle|X^i_t = x\right]. \tag{3.4}$$

Recall that $A^i_{t,T}$ denotes the set of admissible controls for bank i – defined in Section 3.2.2. Note also that each bank is simultaneously solving their own optimization problem, and therefore the value function in (3.4) of bank i may depend on the allocations chosen by other banks within the system. In this sense, the value functions are related and our model's setup can be considered game-theoretic.

Our first result derives the non-local dynamic programming equation (which is often referred to as the Hamilton-Jacobi-Bellman equation) for the value function under regularity.

Proposition 3.1. *If there exist optimal controls and the value function in* (3.4) *is* $\mathcal{C}^{1,1}([0,T),\mathbb{R}_+)$*, then it solves the following non-local partial differential equation (PDE):*

$$\begin{aligned} 0 = \partial_t V_i + \sup_{c_i,w_{i\cdot}} \Bigg\{ &\left[(1-c_i)\,r + \sum_{j\neq i} w_{ij}\mu_j + \frac{\eta_i\mu_i}{\phi_i}\right] x\partial_x V_i + \theta_i \bar{F}_i(c_i)\Big[V_i(t,x(1-\eta_i)) - V_i\Big] \\ &+ \sum_{j\neq i} \theta_j \bar{F}_j(c_j)\Big[V_i(t,x(1-\phi_j w_{ij})) - V_i\Big]\Bigg\} \end{aligned} \tag{3.5}$$

with terminal condition $V_i(T,x) = U_i(x)$*. Where unspecified, the value function and its derivatives are evaluated at* (t,x)*.*

The proof is contained in Appendix 3.A.1, and follows from applying Itô's formula to the value function between t and an appropriately defined sequence of stopping times. Assuming existence of the optimal controls is verified by Corollary 3.3.

Fortunately, it is possible to find a separable solution to (3.5), and explicit solutions for the optimal allocations. It is convenient to introduce the following notation:

$$\Gamma(\delta;\gamma) = \begin{cases} \frac{1-(1-\delta)^{1-\gamma}}{1-\gamma} & \gamma > 0, \gamma \neq 1 \\ -\log(1-\delta) & \gamma = 1, \end{cases} \tag{3.6}$$

for any $\delta \in [0,1)$. Within this range, we note that $\Gamma \geq 0$. There is a natural interpretation of this object; for a utility function of the form in (3.3), $\Gamma(\delta;\gamma)$ is proportional to the loss in utility caused by losing a fraction δ of wealth. More precisely, $\Gamma(\delta;\gamma_i) = x^{\gamma_i - 1}\left[U_i(x) - U_i(x(1-\delta))\right]$ for any $x > 0$.

We can now state our second main result, which presents a solution to (3.5) and computes the optimal allocation of capital.

Proposition 3.2. *The unique optimal cash and interbank lending amounts for the maximization problem in* (3.5) *are given by*

$$\begin{aligned} \hat{c}_i &= \begin{cases} f_i^{-1}\left(\frac{r}{\theta_i \Gamma(\eta_i;\gamma_i)}\right) & \text{if } \frac{r}{\theta_i\Gamma(\eta_i;\gamma_i)} \leq f_i(0) \\ 0 & \text{otherwise} \end{cases} && \forall i = 1, \cdots, n \\ \hat{w}_{ij} &= \begin{cases} \frac{1}{\phi_j}\left(1 - \left(\frac{\phi_j \theta_j \bar{F}_j(\hat{c}_j)}{\mu_j}\right)^{1/\gamma_i}\right) & \text{if } \frac{\phi_j\theta_j\bar{F}_j(\hat{c}_j)}{\mu_j} \leq 1 \\ 0 & \text{otherwise} \end{cases} && \forall j \neq i. \end{aligned} \tag{3.7}$$

Furthermore, with the notation

$$J_i^* = \frac{\eta_i \mu_i}{\phi_i} + (1-\hat{c}_i)r - \theta_i \bar{F}_i(\hat{c}_i)\Gamma(\eta_i;\gamma_i) + \sum_{j \neq i} \hat{w}_{ij}\mu_j - \theta_j \bar{F}_j(\hat{c}_j)\Gamma(\phi_j \hat{w}_{ij};\gamma_i),$$

the following are explicit solutions to (3.5).

(i) if $\gamma_i = 1$ and $U_i(x) = \log x$, we have $V_i(t,x) = g_i(t) + \log x$, where $g_i(t) = (T-t)J_i^*$

(ii) otherwise, for $\gamma_i \neq 1$ and $U_i(x) = \frac{x^{1-\gamma_i}}{1-\gamma_i}$, then we have $V_i(t,x) = g_i(t)U_i(x)$, where $g_i = e^{(1-\gamma_i)(T-t)J_i^*}$.

The proof, which is also given in Appendix 3.A.1, follows from plugging in the proposed solution, simplifying, and then analyzing the necessary and sufficient conditions for optimality of the resulting maximization problem. A key observation in this proof is that the maximization problem in (3.5) is additively separable between each of the controls $c_i, w_{i\cdot}$.

Remark 3.2. *The optimal interbank loan $\hat{w}_{ij}$ depends explicitly on $\hat{c}_j$ through the function $\bar{F}_j(\hat{c}_j)$. Moreover, for any choice of bank j's cash supply, there exists a corresponding optimal value of w_{ij}. In a game-theoretic sense, this would be bank i's best response to j's choice. However, bank j's optimal value $\hat{c}_j$ depends only on fixed model parameters. This ensures that $\hat{c}_j$ is bank j's best response to* any *decisions made by the other banks, and is therefore a dominant strategy. Hence, the 'game' is trivialized – one can compute every other bank' optimal $\hat{c}_j$, after which the corresponding $\hat{w}_{ij}$'s can be easily found.*

The final result of this subsection verifies that the solution given in Proposition 3.2 is indeed equal to the value function.

Corollary 3.3. *The value function in* (3.4) *is given by*

$$V_i(t,x) = \begin{cases} g_i(t) + \log x & \text{if } \gamma_i = 1 \\ g_i(t)\frac{x^{1-\gamma_i}}{1-\gamma_i} & \text{otherwise,} \end{cases}$$

where $g_i(t)$ and the optimal controls are given in Prop. 3.2.

The proof in Appendix 3.A.1 uses a verification argument. We show that any solution to (3.5) that is once continuously differentiable in both time and space is equal to the value function. Since the proposed solutions in Proposition 3.2 satisfy this regularity condition, we conclude the desired claim. Finally, this result verifies the assumption made in Proposition 3.1 regarding the existence of optimal controls.

Analysis of Decentralized Optimum

With explicit solutions for the optimal allocations, it is possible to analyze their dependence on the exogenous parameters of the system. First, note that the optimal interbank loan $\hat{w}_{ij}$ depends on bank i only through their risk aversion parameter γ_i. Hence, if $\gamma_i = \gamma_k$ then $\hat{w}_{ij} = \hat{w}_{kj}$. Although the fractional amount of these interbank loans are equal, the nominal amounts may differ. However, the optimal lending amount is decreasing in the lender's risk aversion coefficient γ_i, as we might expect.

From (3.7), we can also see that $\hat{c}_i$ is decreasing in the risk-free rate. This occurs because cash is effectively depreciating at the risk-free rate r. However, each unit of additional cash provides a marginal benefit by lowering the risk of a bank's project failing. From the proof of Proposition 3.2, the optimal choice of $\hat{c}_i$ will solve the following:

$$\max_{c_i \geq 0} \{ -rc_i - \theta_i \bar{F}_i(c_i)\Gamma(\eta_i;\gamma_i)\},$$

which indicates that the resulting $\hat{c}_i$ achieves the optimal tradeoff between the cost of liquidity and induced risk. In particular, the optimal $\hat{c}_i$ ensures that the marginal cost of holding liquidity (r) equals the marginal benefit of reducing risk ($\theta_i f_i(\hat{c}_i)\Gamma(\eta_i;\gamma_i)$). In the extreme case where r is large, it may be too costly (relative to the potential losses) for a bank to hold any amount of cash, i.e. $\hat{c}_i = 0$.

The quantity $\frac{\mu_j}{\phi_j \theta_j \bar{F}_j(\hat{c}_j)}$, which appears in (3.7) for $\hat{w}_{ij}$, is similar to the well-known Sharpe ratio. However, there is one main difference. The variance of returns for bank j's project can be controlled by bank j itself. Nonetheless, notice that the optimal investment $\hat{w}_{ij}$ grows with this 'Sharpe-like' ratio. If, in particular, the ratio is less than one, then the expected excess return of the interbank loan (equal to $\mu_j - \phi_i \theta_i \bar{F}_j(\hat{c}_j)$) is negative, and bank i would in fact prefer to short project j. Since this is not permitted in our model, bank i resorts to an investment of zero. As a direct result, notice that network's sparsity structure is dictated by this quantity – a bank j has creditors if and only if $\frac{\mu_j}{\phi_j \theta_j \bar{F}_j(\hat{c}_j)} > 1$. This implies a 'core-periphery' network structure, such that a subset of banks serve as the only borrowers – an example of such a financial network can be seen in Figure 3.

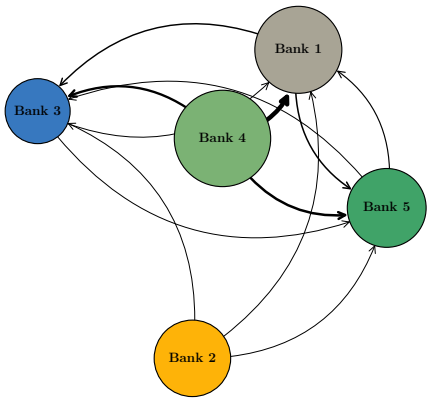


Figure 3: Sample financial network generated by the decentralized optimum. Edges point from lending to borrowing banks, and their width indicates the nominal size of the exposures. Node size indicates total capitalization.

Bank	μ_i (%)	ϕ_i	η_i	θ_i	γ_i	$\bar{F}_i(x)$
1	0.9	0.2	0.5	0.04	0.5	$e^{-0.5x}$
2	1	0.3	0.6	0.08	1.7	$e^{-0.6x}$
3	1.5	0.9	0.7	0.12	1	$e^{-0.7x}$
4	1.3	0.6	0.4	0.05	0.3	e^{-2x}
5	1.3	0.82	0.9	0.02	0.87	$e^{-2.4x}$

Table 1: Parameters for the financial network in Figure 3. The risk-free rate is equal to $r = 5\%$. Code generating this figure can be found here.

3.3.2 Centralized Network

Consider now the perspective of a single central planner of the financial system. In contrast with Section 3.3.1, we will see that the planner has two different incentives for bank i's holding of cash.

The first is identical – bank i stands to lose wealth if their project fails. The second incentive is systemic – other banks face losses on their interbank claims upon the very same event. Therefore, we expect the planner to have stronger incentives to hold cash, and elect for a greater supply of liquidity within the system.

We assume that the planner seeks to maximize the total welfare in the system – defined as the sum of all banks' utilities. Their value function is therefore given by the following:

$$V(t, x_1, ..., x_n) = \sup_{(c_\cdot, w_\cdot) \in \mathcal{A}_{t,T}} \mathbb{E}\left[\sum_{i=1}^{n} U_i(X_T^i) \middle| (X_t^1, \dots X_t^n) = (x_1, \dots x_n) \right], \tag{3.8}$$

where $\mathcal{A}_{t,T} = \prod_i \mathcal{A}_{t,T}^i$ is the Cartesian product of each bank's admissible controls.

Remark 3.3. *It is important to note that there are many possible 'social welfare functions' for the planner to consider. In this section, we will see that using the sum of utilities allows for separable solutions to the value function when all banks have logarithmic utility, i.e. $\gamma_i = 1$ for all i. We note that if the planner maximized the* product *of utilities, then we can also find an explicit solution and optimal controls in the case where $\gamma_i \in (0, 1)$ for all i, but we omit these calculations for conciseness.*

Notice that we can relate the planner's value function to those of individual banks from (3.4). The optimal decentralized allocation from Section 3.3.1 is always feasible for the planner, and therefore their value function is bounded from below by the sum of each bank's value function as follows:

$$V(t, x_1, ..., x_n) \geq \sum_{i=1}^{n} V_i(t, x_i). \tag{3.9}$$

This inequality reflects an inefficiency of the decentralized setting; the planner's allocation is the first-best (i.e. welfare-maximizing) outcome for the system. In what follows, we analyze the planner's optimal allocation by deriving the dynamic programming equation and analyzing the resulting optimization problem. As in the previous section, we first derive the non-local (PDE) solved by the planner's value function.

Proposition 3.4. *If there exist optimal controls, and the value function in* (3.8) *is $\mathcal{C}^{1,1,\dots,1}([0,T), \mathbb{R}_+, ..., \mathbb{R}_+)$, then it solves*

$$\begin{aligned} 0 = \partial_t V + \sup_{c_\cdot, w_\cdot} \Bigg\{ \sum_{i=1}^{n} \Bigg(\Bigg[(1 - c_i)\, r + \sum_{j \neq i} w_{ij} \mu_j + \frac{\eta_i \mu_i}{\phi_i} \Bigg] x_i \partial_{x_i} V \\ + \theta_i \bar{F}_i(c_i) \Big[V(t, x_1(1 - \phi_i w_{1i}), .., x_i(1 - \eta_i), .., x_n(1 - \phi_i w_{ni})) - V \Big] \Bigg) \Bigg\}, \end{aligned} \tag{3.10}$$

with terminal condition $V(T, x_1, .., x_n) = \sum_{i=1}^{n} U_i(x_i)$. *Where unspecified, the value function and its derivatives are evaluated at* $(t, x_1, ..., x_n)$.

The proof of this result is only a minor adaptation to the proof of Proposition 3.1, and can be found in Appendix 3.A.2.

Proposition 3.4 yields an $n+1$ dimensional non-local PDE for the planner's value function. There is one key difference between Equations (3.10) and (3.5) – when a project fails, the planner's value function is affected by losses occurring throughout the entire financial system. This is not true in the decentralized setting; an individual bank's value function only depends on their own losses incurred by such a failure.

With specific choices of utility functions, it is possible to find a separable solution to (3.10), and prove existence of an optimal allocation. However, to establish uniqueness, we will need the following technical assumption.

Assumption 3.4. *Let each shock density* $f_i(\cdot)$ *satisfy*

$$\frac{f_i(c)}{\bar{F}_i(c)} + 3\frac{f_i'(c)}{f_i(c)} - \frac{f_i''(c)}{f_i'(c)} < 0, \ \forall c \geq 0. \tag{3.11}$$

and, with the notation $\tilde{c}_i = F_i^{-1}\left(\left[1 - \frac{\mu_i}{\phi_i\theta_i}\right]_+\right)$, *assume that the following holds for all* i

$$\Gamma(\eta_i; 1) > \begin{cases} \min\left\{(n-1)\left[\log\left(\frac{\phi_i\theta_i}{\mu_i}\right) - \frac{f_i(0)^2}{f_i'(0)}\right], \ \frac{r}{\theta_i f_i(0)} + (n-1)\log\left(\frac{\phi_i\theta_i}{\mu_i}\right)\right\} & \text{if } \tilde{c}_i = 0 \\ \min\left\{-(n-1)\frac{\phi_i\theta_i f_i(\tilde{c}_i)^2}{\mu_i f_i'(\tilde{c}_i)}, \ \frac{r}{\theta_i f_i(\tilde{c}_i)}\right\} & \text{otherwise.} \end{cases} \tag{3.12}$$

Assumption 3.4 is sufficient for uniqueness of the planner's the optimal allocation. While we have numerically observed that the optimal solution is almost always unique, the optimization problem in (3.10) is (generally) not convex, and therefore proving uniqueness is non-trivial. We do, however, note that the inequality (3.11) is always satisfied by exponential and power distributions.

Analogous to Section 3.3.1, we show there exists a separable solution to the PDE (3.10). Additionally, we show that the optimal solution will solve a system of algebraic equations.

Proposition 3.5. *Let each bank have a logarithmic utility function (i.e.* $\gamma_i = 1 \ \forall i$*). Then, there exist optimal cash and lending amounts for the planner, which solve the following system of equations:*

$$
\begin{aligned}
c_i^* &= \begin{cases} f_i^{-1}\left(\frac{r}{\theta_i\left[\Gamma(\eta_i;1)+(n-1)\Gamma(\phi_i w_{\cdot i}^*;1)\right]}\right) & \text{if } f_i(0) \leq \frac{r}{\theta_i\left[\Gamma(\eta_i;1)+(n-1)\Gamma(\phi_i w_{\cdot i}^*;1)\right]} \\ 0 & \text{otherwise,} \end{cases} \\
w_{\cdot i}^* &= \begin{cases} \frac{1}{\phi_i}\left(1-\frac{\phi_i\theta_i\bar{F}_i(c_i^*)}{\mu_i}\right) & \text{if } \frac{\phi_i\theta_i\bar{F}_i(c_i^*)}{\mu_i} \leq 1 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.13}
$$

Letting c_i^* and $w_{\cdot i}^*$ be the optimal allocation, we define

$$
J_C^* = \sum_{i=1}^{n}\left(\left[(1-c_i^*)\,r + (n-1)w_{\cdot i}^*\mu_i + \frac{\eta_i\mu_i}{\phi_i}\right] - \theta_i\bar{F}_i(c_i^*)\Big[\Gamma(\eta_i;1)+(n-1)\Gamma(\phi_i w_{\cdot i}^*;1)\Big]\right),
$$

and $g(t) = (T-t)J_C^*$. The solution to (3.10) *is given by*

$$
V(t,x_1,\ldots,x_n) = g(t) + \sum_{i=1}^{n}\log x_i. \tag{3.14}
$$

Furthermore, under Assumption 3.4, the optimal cash and lending amounts $(c_i^*, w_{\cdot i}^*)$ are unique.

The proof is again given in Appendix 3.A.2. We note that a separable solution using logarithmic utility functions is only possible because the planner aims to maximize the sum of expected utilities. See Remark 3.3 for a brief discussion of other settings where a separable solution can be obtained.

In contrast to the decentralized setting, the maximization in (3.10) is not additively separable between each optimization variable. Nonetheless, each of the i subsets $\{c_i, w_{1i}...w_{ni}\},\ i = 1...n$ can be analyzed separately, which greatly simplifies our analysis. However, the coupling between c_i and $w_{\cdot i}$ leads to the need for additional assumptions to establish uniqueness.

The system of equations in (3.13) admits a block coordinate descent approach. Namely, for any fixed c_i, the maximization problem for $w_{\cdot i}$ is strictly concave and admits a unique solution (these can be seen in the proof of Proposition 3.5). Conversely, for given values of $w_{\cdot i}$, the maximizing of c_i shares these features. As a result, we can iteratively update these variables to solve for the planner's optimum numerically. Upon convergence, we are guaranteed to have found the unique optimal allocation.

Since we have shown existence of an optimal allocation, and the proposed solution in (3.14) is continuously differentiable, then we are able to verify that it is indeed equal to the planner's value function.

Corollary 3.6. *The planner's value function in* (3.8) *is given by* (3.14)*. Furthermore, the optimal*

interbank lending and cash amounts solve (3.7).

Analysis of Centralized Optimum

There is one main difference between the system of equations in (3.13) and the optimal solutions in (3.7) obtained from the decentralized setting. Here, we have an additional term of $(n-1)\Gamma(\phi_i w^*_{\cdot i}; 1)$ that influences the planner's optimal value for c^*_i. This term directly captures the externality – when i's project fails, the planner sees losses in utility experienced by all banks. As a result, with more banks the planner maintains a larger supply of cash to compensate for greater systemic losses. In contrast, bank i's decentralized optimization problem considers only changes to their own wealth, and therefore their optimal $\hat{c}_i$ will be indifferent to the system's size.

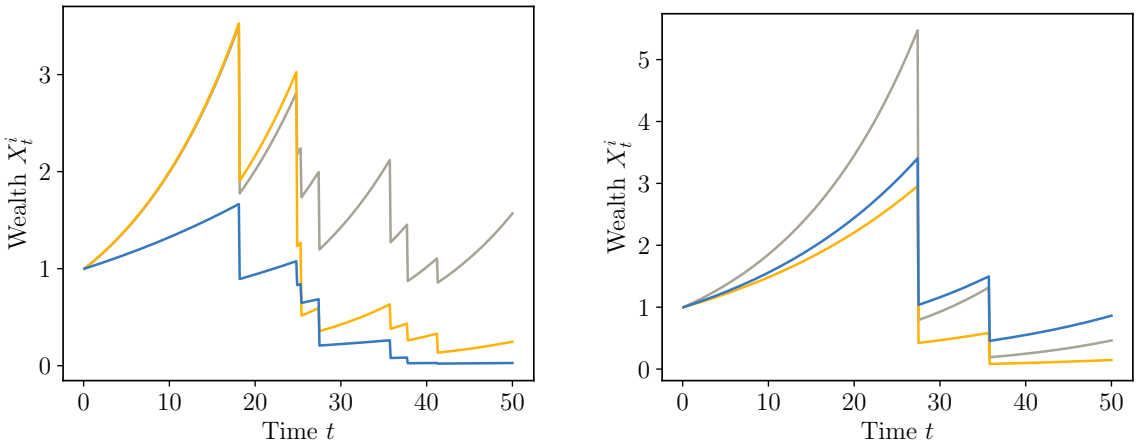
Since we will have $w^*_{\cdot i} \geq 0$ in (3.13), the planner has a greater incentive to hold liquidity than the individual bank.[5] Hence, the planner will hold more liquidity than the decentralized optimal allocation – we will study this difference more closely in the following section. Finally, we also notice that given the amounts of cash held, the optimal investments $w^*_{\cdot\cdot}$ and $\hat{w}_{\cdot\cdot}$ are computed identically. It follows that any differences between the optimal interbank lending amounts in (3.7) and (3.13) can only be driven by differences in optimal cash supplies.

3.4 Price of Anarchy

It is natural to compare the two optimal allocations from Sections 3.3.1 and 3.3.2. In particular, we may be interested in computing the gap in welfare from the inequality (3.9). More generally, in simulations we see stark differences between the two optimal allocations. Figure 4 illustrates a sample path for the wealth of three banks, where in 4a the controls are given by (3.7), and in 4b by (3.13). Qualitatively, there are higher-frequency jumps in 4a, but the jumps are of larger size in 4b. With the remainder of this section, we study these differences more precisely.

In what follows, we will assume that all banks have logarithmic utility (i.e. $\gamma_i = 1$ for all i). Recall that $\hat{c}_i, \hat{w}_{ji}$ denote the optimal decentralized allocations given in (3.7). Note that for all $j, k \neq i$ we will have $\hat{w}_{ki} = \hat{w}_{ji}$, so we will denote these fractional amounts to be $\hat{w}_{\cdot i}$ (this follows from $\gamma_j = 1$ for all j). Additionally, recall that $c^*_i, w^*_{\cdot i}$ denotes the optimal solution from (3.13). Finally, we use the asymptotic notation $g(n) = \Theta(h(n))$ to denote that there exist positive constants A_1, A_2 such that $A_1 \leq \lim_{n\to\infty} \frac{g(n)}{h(n)} \leq A_2$. If $A_1 = A_2$, then we will write $g(n) \asymp h(n)$.

[5]This observation may not be the case if, for example, short-selling were allowed. Qualitatively, the planner may choose to have a single bank i hold zero cash, while others in the system maintain large, short positions in i's project. In this case, the total utility of the system may actually increase when bank i's project fails. However, clearly this result may not align with the best outcome for bank i itself.



(a) Decentralized

(b) Centralized

Figure 4: An example of wealth dynamics under the both optimal allocations for a system of $n = 3$ banks. The same random seed is used in both simulations, so that the size and arrival times of liquidity shocks are identical. For conciseness, we do not include the parameters, but the code to reproduce these figures can be found here.

3.4.1 Liquidity Supply and Project Risk

Comparing the two optimal allocations, since $w^*_{\cdot i} \geq 0$, it will necessarily be the case that $c^*_i \geq \hat{c}_i$. Our fundamental result establishes the asymptotic rate at which the planner's optimal supply of liquidity grows as the size of the system increases. More precisely, we show that for heavy-tailed distributions, the planner's supply of cash must grow at least logarithmically in the system size n – and under stronger assumptions this lower bound is tight. In contrast, if $w^*_{\cdot i} = 0$, then we would have $c^*_i = \hat{c}_i$ – which is of constant order.

Proposition 3.7. *Assume that $w^*_{\cdot i} > 0$. If the shock density satisfies: $f_i(x) \geq \kappa_{i,L} e^{-\frac{x}{\lambda_{i,L}}}$, for all x and fixed constants $\lambda_{i,L} > 0$ and $\kappa_{i,L} > 0$, then*

$$c^*_i \geq \lambda_{i,L} \log\left(\frac{\theta_i \kappa_{i,L} \Gamma(\phi_i \hat{w}_{\cdot i}; 1)}{r}\right) + \lambda_{i,L} \log(n-1).$$

In particular, the planner's optimal cash supply asymptotically grows at least logarithmically in n.

Furthermore, if for all x we also have:

$$f_i(x) \leq \kappa_{i,U} e^{-\frac{x}{\lambda_{i,U}}}$$

for $\lambda_{i,L} \leq \lambda_{i,U}$ and $\kappa_{i,L} \leq \kappa_{i,U}$, then

(i) ***Upper Bound:***

$$c_i^* \leq \lambda_{i,U} \log\left(\frac{\theta_i \kappa_{i,U} C_U}{r}\right) + \lambda_{i,U} \log\left((n-1)\log(n)\right),$$

where $C_U > 3$ depends on all model parameters (including $\lambda_{i,L}$ and $\lambda_{i,U}$), but does not explicitly grow with n. As a result, $\lim_{n\to\infty} \frac{c_i^*}{\log(n)} \leq \lambda_{i,U}$.

(ii) ***Lower Bound:***

$$c_i^* \geq \lambda_{i,L} \log\left(\frac{\theta_i \kappa_{i,L} \lambda_{i,L}}{r \lambda_{i,U}}\right) + \lambda_{i,L} \log\left((n-1)\Big[\log(n-1) - \frac{\lambda_{i,U}}{\lambda_{i,L}} \log\left(C_L\right)\Big]\right),$$

for $C_L > 0$ depending only on i's parameters. Hence, $\lim_{n\to\infty} \frac{c_i^*}{\log(n)} \geq \lambda_{i,L}$.

Combining the two limiting bounds, we have $c_i^* = \Theta\big(\log(n)\big)$.

The proof is provided in Appendix 3.A.3. It follows from iterating through upper (and lower) bounds for c_i^* using the system of equations in (3.13), and beginning from crude estimates.[6]

The following special case of Proposition 3.7 occurs when the shock sizes are exponentially distributed.

Corollary 3.8. *If $F_i(x) = 1 - e^{-\frac{x}{\lambda_i}}$, then*

$$\lambda_i \log\left(\frac{\theta_i(n-1)}{\lambda_i r}\left[\log(n-1) - \log\left(\frac{\Gamma(\eta_i;1)}{\Gamma(\phi_i \hat{w}_{\cdot i};1)}\right)\right]\right) \leq c_i^* \leq \lambda_i \log\left(\frac{\theta_i C_U (n-1)}{\lambda_i r}\log(n)\right).$$

In particular, $c_i^* \asymp \lambda_i \log(n)$.

The proof follows from plugging in $\lambda_{i,L} = \lambda_{i,U} = \lambda_i$, and $\kappa_{i,L} = \kappa_{i,U} = \lambda_i^{-1}$. Simplifying the constant C_L that appears in the lower bound yields the desired result.

Corollary 3.8 is a useful tool for comparing the two optimal allocations, as all differences are driven by the distinct supply of liquidity. From here onward, we assume the setting of this Corollary, wherein all shock sizes are exponentially distributed. First, we can directly compute the dependence of project i's likelihood of failure on the system size. We see that:

$$\frac{r\lambda_i}{C_U(n-1)\log(n)} \leq \theta_i \bar{F}_i(c_i^*) \leq \frac{r\lambda_i}{(n-1)\left[\log(n-1) - \log\left(\frac{\Gamma(\eta_i;1)}{\Gamma(\phi_i \hat{w}_{\cdot i};1)}\right)\right]},$$

[6] It is possible to use the same techniques in this proof to obtain bounds when the density has power-law tails. While the results are not qualitatively different, we are unable to achieve the tight bound that appears in Corollary 3.8 when the shock distribution is itself a power-law. The main result can be seen in Appendix 3.B.

and it follows that $\bar{F}_i(c_i^*) = \Theta\left(\frac{1}{n\log(n)}\right)$.[7] In stark contrast, the optimal intensity from the decentralized setting, $\bar{F}_i(\hat{c}_i)$, is constant in n. That is, $\bar{F}_i(\hat{c}_i) = \Theta(1)$. These two results will allow us to analyze the price of anarchy.

3.4.2 Losses to Lending Banks

In addition to the supply of liquidity, the optimal investment amounts will differ between the two settings. Due to the greater risk of jumps in the decentralized optimum, banks will invest less capital into each others' projects, and have a lesser degree of integration with the system. Hence, we are also interested in comparing the losses experienced by lenders when i's project fails.

First, we study the asymptotics of $w_{\cdot i}^*$. Recalling that $f_i(\cdot)$ is assumed to be exponential, having already shown that the intensity $\bar{F}_i(c_i^*)$ is of asymptotic order $\frac{1}{n\log(n)}$, (3.13) allows us to easily compute:

$$w_{\cdot i}^* = \frac{1}{\phi_i} - \Theta\left(\frac{1}{n\log(n)}\right).$$

Note that we must have $w_{\cdot i} < \phi_i^{-1}$ to ensure wealth remains positive, yet we can still pin down the rate at which the interbank investment approaches its upper bound.

Next, we are interested in the term $\Gamma(\phi_i w_{\cdot i}^*; 1) = -\log\left(1 - \phi_i w_{\cdot i}^*\right)$, which represents the relative loss of utility to a single lender when bank i's project experiences a failure. A straightforward computation using (3.6) gives

$$\Gamma(\phi_i w_{\cdot i}^*; 1) = \Theta\big(\log(n\log(n))\big) = \Theta\big(\log(n)\big),$$

since $\frac{\log(n\log(n))}{\log(n)} \underset{n\to\infty}{\to} 1$.

Putting this result together with the asymptotic rate of $\bar{F}_i(c_i^*)$, we see that

$$\bar{F}_i(c_i^*)(n-1)\Gamma(w_{\cdot i}^*; 1) = \Theta(1).$$

This is an interesting result, as it shows that the expected losses of utility due to a project's failure do not grow with the system size – in contrast, the decentralized setting exhibits $\bar{F}_i(\hat{c}_i)(n-1)\Gamma(\hat{w}_{\cdot i}; 1) = \Theta(n)$. Namely, the planner perfectly compensates for larger expected losses in utility through its reduction of a project's failure probability.

[7] We can obtain similar bounds using only Proposition 3.7, but these will not be tight. In particular, we would only show that $\bar{F}_i(c_i^*) = O\left(\left(n\log(n)\right)^{-\frac{\lambda_{i,L}}{\lambda_{i,U}}}\right)$, and $\bar{F}_i(c_i^*) = \Omega\left(\left(n\log(n)\right)^{-\frac{\lambda_{i,U}}{\lambda_{i,L}}}\right)$.

3.4.3 Price of Anarchy Asymptotics

We now turn to the gap between value functions from (3.9). It will be useful to have $\mathcal{M}_n$ denote the set of banks that are lent a non-zero amount of capital in the planner's optimal allocation, i.e. $\mathcal{M}_n = \{i \in [1..n] : w^*_{\cdot i} > 0\}$. Banks in $\mathcal{M}_n$ form the 'core' of the financial network. If for some i we have $w^*_{\cdot i} = 0$, then it must be the case that $c^*_i = \hat{c}_i$ and $\frac{\phi_i \theta_i \bar{F}_i(c^*_i)}{\mu_i} > 1$. For such a bank i, the planner's optimal c^*_i would remain constant at $\hat{c}_i$, even as n grows.

The 'price of anarchy' reflects how greedy decentralized behavior leads to lesser welfare in the system (Papadimitriou, 2001). In this model, we define it as

$$\text{PoA} = \frac{V}{\sum_{i=1}^n V_i}.$$

More precisely, the price of anarchy equals the relative loss in value between the centralized and decentralized settings. In the following result, we characterize its asymptotic behavior.

Proposition 3.9. *Assume that $\gamma_i = 1$ and $F_i(x) = 1 - e^{-\frac{x}{\lambda_i}}$ for all i. Then, as $n \to \infty$, we have*

$$\text{PoA} = \Theta(1).$$

The proof is found in Appendix 3.A.3, and uses all previous results from this Section.

It is particularly interesting that the price of anarchy does not grow with the system size n, or the remaining time horizon $(T - t)$. A more precise result can be obtained if banks are sufficiently homogeneous, where we can compute the limiting price of anarchy.

Corollary 3.10. *Assume that all banks in $\mathcal{M}_n$ are identical (i.e. $\mu_j = \mu$, $\phi_j = \phi$, $\theta_j = \theta$, $\eta_j = \eta$, and $\lambda_j = \lambda$ for some given constants μ, ϕ, θ, η and λ). If $|\mathcal{M}_n| \underset{n\to\infty}{\to} \infty$, then*

$$\begin{aligned}
&\frac{V_i}{|\mathcal{M}_n|(T-t)} \underset{n\to\infty}{\to} \frac{\mu}{\phi} + \theta \bar{F}(\hat{c})) \left[\log\left(\frac{\phi\theta\bar{F}(\hat{c})}{\mu}\right) - 1\right], \quad \forall i = 1...n \\
&\frac{V}{n|\mathcal{M}_n|(T-t)} \underset{n\to\infty}{\to} \frac{\mu}{\phi}.
\end{aligned}$$

where $\hat{c}$ is given in (3.7) *and $\bar{F}(\hat{c}) = e^{-\frac{\hat{c}}{\lambda}}$. As a result, we have:*

$$\text{PoA} \underset{n\to\infty}{\to} \frac{1}{1 + \frac{\phi\theta\bar{F}(\hat{c})}{\mu}\left[\log\left(\frac{\phi\theta\bar{F}(\hat{c})}{\mu}\right) - 1\right]}. \tag{3.15}$$

Corollary 3.10 verifies that the price of anarchy is of constant order n, and the proof is found in Appendix 3.A.3. Of particular interest, the rate at which $|\mathcal{M}_n|$ grows in n does not appear in our result. This implies that the limiting price of anarchy is independent from the fraction of the system that operates as its 'core'. Notice also that $\phi\theta\bar{F}(\hat{c}) < \mu$, and hence the right-hand side in (3.15) is greater than one. Moreover, the limiting price of anarchy is increasing in $\frac{\phi\theta\bar{F}(\hat{c})}{\mu}$. Therefore, as the profitability of interbank loans in the decentralized setting is reduced, the limiting price of anarchy grows to infinity.

Corollary 3.10 is verified numerically. Using the parameters in Table 2, we compute the individual and collective value functions. The price of anarchy is plotted in Figure 5, along with the limiting value in (3.15). We see that the price of anarchy quickly converges to the limit.

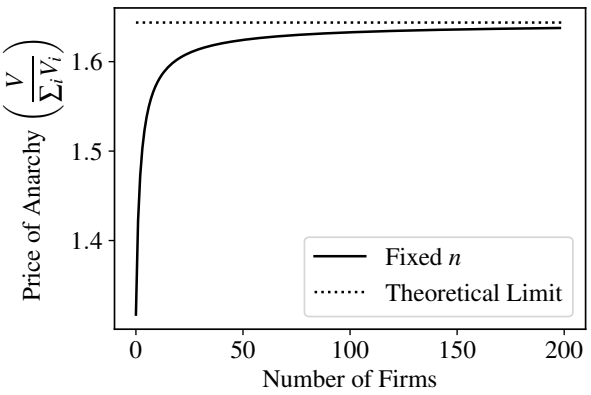


Figure 5: Simulating the Price of Anarchy for a system of identical firms as n grows.

Notation	Value	Description
r	0.01	Risk-free rate
μ	0.045	Excess drift
ϕ	0.4	Losses to lenders
η	0.5	Losses to borrower
$F(x)$	$1 - e^{-\lambda x}$	CDF of shock size
λ	1	Parameter of $F(\cdot)$
θ	0.1	Shock arrival rate

Table 2: System parameters used in simulations for Figure 5. Code is available here.

3.4.4 Replicating the Centralized Allocation

Finally, we may be interested in studying how banks in the decentralized setting can be incentivized to replicate the planner's optimal allocation. To do so, we will allow the degree of each bank's investment into their own project, η_i, to vary. The rationale for this is twofold: first, η_i plays the fundamental role in decentralized banks' choice of how much cash to hold. If this value is sufficiently large, banks will increase their supply of liquidity and therefore reduce their projects' riskiness – which can lead them to meet the planner's optimal allocation. Second, we can imagine that lending banks are permitted to write a contract stipulating the borrowing bank's degree of co-investment. This kind of contracting is not a focus of our work, and is instead analyzed in more detail with Principal-Agent problems such as Hernández Santibáñez et al. (2020). Nonetheless, the co-investment contract can be designed to ensure that individual banks hold sufficient liquidity.

Let η_i^C (resp. η_i^D) denote the fraction of bank i's wealth lost upon project failure in the centralized (resp. decentralized) setting. We would like to choose η_i^D so that decentralized banks replicate the centralized optimum with values η_i^C. More precisely, we seek to find η_i^D solving $c_i^*(\eta_i^C) = \hat{c}_i(\eta_i^D)$ for all i, where we write the optimal controls in a way that highlights their dependence on the underlying values of η. Even though the optimal allocations are identical, however, we note that the decentralized optimum is still inefficient (with respect to the optimal centralized allocation corresponding to η_i^D). Using equations (3.7) and (3.13), we find that:

$$\eta_i^D = 1 - (1 - \eta_i^C)\left(1 - \phi_i w_{\cdot i}^*(\eta_i^C)\right)^{n-1}.$$

First, notice that whenever $w_{\cdot i}^*(\eta_i^C) > 0$, the resulting value of η_i^D will grow exponentially in n towards its upper bound of 1. This is intuitive – banks whose projects are highly invested in require the strongest incentive to reduce their project's risk. Second, we see that for banks to replicate the planner's optimum, it is necessary for bank i's degree of co-investment to depend on their liabilities throughout the system. It is therefore necessary to know the complete structure of the financial network to determine the value of η_i^D, which may not be known to individual lenders. Finally, an interesting case occurs when we choose $\eta_i^C = 0$. In this case, the value of η_i^D is only non-zero if $w_{\cdot i}^*(0) > 0$. Namely, banks without counterparties hold no stake in their own projects.

3.5 Discussion and Conclusion

In this chapter, we present a model by which banks in a financial system control both their own levels of risk, and their investment in each others' risky projects.

We compute the uniquely optimal allocations of capital for two distinct organizations of the system, and study their differences qualitatively and quantitatively. First, we analyze the setting where each bank acts with pure self-interest. We compute explicitly the optimal allocation, and find that the size of interbank investments are closely related to a Sharpe-like ratio – which is controlled by borrowing banks. In particular, the optimal financial network exhibits a 'core-periphery' structure, wherein only a subset of banks serve as borrowers. Second, we formulate the optimization problem of a central planner, who seeks to maximize the total welfare in the system. Under a few technical assumptions, we are able to prove the existence of a unique optimal allocation. In particular, we find that the planner's optimum exhibits low-frequency and high-severity events of distress, which aligns with the 'robust-yet-fragile' feature observed by Gai and Kapadia (2010). The difference in these

two optimal allocations is driven by a negative externality, where individual banks are excessively risky given the potential losses that they may induce.

In the case where shocks are exponentially distributed, we can precisely compute how the externality's severity depends on the system's size. We see that the planner compensates for an increased number of counterparties by reducing the risk of a bank's project. The planner perfectly balances the two effects, so that the expected losses in utility remain of constant order – regardless of the system size. We are also able to see that the loss in welfare due to decentralized behavior grows with the size of both the financial system and its core. However, and perhaps counterintuitively, the *relative* loss of welfare, which we refer to as the price of anarchy, is of constant order. Finally, we show that it is possible, through regulation or contracting between banks, to replicate the planner's optimal interbank allocation. Banks who have borrowed the largest amount of capital will be subjected to the strictest requirements, and will therefore have the strongest incentive to reduce their project's riskiness. This highlights the danger of government bailouts, which can cause perverse incentives for individual banks.

We believe there are several interesting continuations of this work. First, a notable limitation of this model is that it does not contain a mechanism of contagion. For instance, Aït-Sahalia and Hurd (2015) consider a portfolio optimization problem where assets' jump components are self- and mutually exciting. An immediate extension of our work may be to incorporate jump processes with these features directly into the model. It may also be possible to show that self- and mutually exciting jumps can endogenously emerge, e.g. if a lending bank suffers losses of liquidity when their borrowers' project fails. Additionally, financial crises are heavily destabilizing, and it is natural to assume that it is challenging (or impossible) to quickly rebalance a portfolio in the wake of such an event. Therefore, it is practical to prevent banks from instantaneously re-weighting their portfolios. This feature may lead to further inefficiencies caused by banks' inability to establish an optimal allocation of wealth shortly after a shock occurs. Furthermore, our model differs from the literature on strategic network formation in that creating a 'lending linkage' to another bank is costless. It is natural to incorporate these costs into banks' optimization problems, for example, as the cost of performing due diligence on a borrower to assess their creditworthiness. Finally, the inclusion of intermediary costs or more sophisticated contracting mechanisms between banks presents a rich direction of future research.

Appendices

3.A Proofs

3.A.1 Decentralized Network

Proof of Proposition 3.1. First, we use the dynamic programming principle to consider only the optimal control over the time interval $[t, \tau]$, for a stopping time $\tau < T$ to be defined later. We can write the value function recursively as

$$V_i(t,x) = \sup_{(c^i_\cdot, w^{i\cdot}_\cdot) \in \mathcal{A}^i_{t,T}} \mathbb{E}\Big[V_i\left(\tau, X^i_\tau\right) \Big| X^i_t = x\Big], \tag{3.16}$$

which holds for all $t < T$ and $\tau \leq T$.

Next, we for each bank k we fix some admissible control $(c^k_\cdot, w^{k\cdot}_\cdot) \in \mathcal{A}^k_{t,T}$. By assumption, V_i is once differentiable in both time and space, and using Itô's formula (see for instance Cont and Tankov (2003)) we can write:

$$\begin{aligned} V_i\left(\tau, X^i_\tau\right) - V_i\left(t, X^i_t\right) = & \int_t^\tau \left[\partial_t V_i(s, X^i_s) + \partial_x V_i\left(s, X^i_s\right) b_i(c^i_s, w^{i\cdot}_s) X^i_s\right] ds \\ & + \sum_{j=1}^n \int_t^\tau \left[V_i\left(s, X^i_s\right) - V_i\left(s, X^i_{s-}\right)\right] dN^j_s. \end{aligned} \tag{3.17}$$

where $b_i(c^i_t, w^{i\cdot}_t)$ is the coefficient on the dt term in (3.2).

Recall that the jump process N^j_t has instantaneous intensity $\theta_j \bar{F}_j(c^j_t)$. Therefore, the compensated process $M^j_t = N^j_t - \int_0^t \theta_j \bar{F}_j(c^j_s) ds$ is a martingale. Rewriting the integrals in (3.17) in terms of dM^j_t and taking expectation conditioned on $X^i_t = x$ (denoted $\mathbb{E}_{t,x}$) of both sides yields:

$$\begin{aligned}\mathbb{E}_{t,x}\left[V_i\left(\tau, X_\tau^i\right)\right] - V_i\left(t, X_t^i\right) = \mathbb{E}_{t,x}&\left[\int_t^\tau \mathcal{L}^{c_s^i, w_s^{i\cdot}} V_i(s, X_{s-}^i) ds\right] \\ &+ \mathbb{E}_{t,x}\left[\int_t^\tau \left[V_i\left(s, X_{s-}^i - \eta_i X_{s-}^i\right) - V_i\left(s, X_{s-}^i\right)\right] dM_s^i\right] \\ &+ \sum_{j\neq i} \mathbb{E}_{t,x}\left[\int_t^\tau \left[V_i\left(s, X_{s-}^i - \phi_j w_s^{ij} X_{s-}^i\right) - V_i\left(s, X_{s-}^i\right)\right] dM_s^i\right],\end{aligned} \tag{3.18}$$

where the generator $\mathcal{L}^{c_i, w_{i\cdot}}$ is defined to be

$$\begin{aligned}\mathcal{L}^{c_i, w_{i\cdot}}\psi(t,x) = \partial_t\psi(t,x) &+ \left((1-c_i)r + \sum_{j\neq i} w_{ij}\mu_j + \frac{\eta_i\mu_i}{\phi_i}\right) x\partial_x\psi \\ &+ \theta_i\left(1 - F_i(c_i)\right)\Big[\psi(t, x(1-\eta_i)) - \psi(t,x)\Big] \\ &+ \sum_{j\neq i}\theta_j\left(1 - F_j(c_j)\right)\Big[\psi(t, x(1-\phi_j w_{ij})) - \psi(t,x)\Big],\end{aligned} \tag{3.19}$$

for any $\psi \in \mathcal{C}^{1,1}([0,T), \mathbb{R}_+)$.

Next, we need to show that the expectation of the stochastic integrals with respect to dM_s^k are equal to zero. To do so, it is sufficient to have the integrand bounded for $s \in [t, \tau]$. Define the stopping time τ to be:

$$\tau = (t+\delta) \wedge \inf\left\{s \in [t,T], X_s^i \leq \epsilon \text{ or } X_s^i \geq \frac{1}{\epsilon}\right\}, \tag{3.20}$$

for some small $\delta > 0$ and $\epsilon > 0$. Then, since X_s^i is bounded away from zero in $[t, \tau]$, the size in the jump of the value function is bounded. Therefore the stochastic integrals in (3.18) have zero expectation. We obtain:

$$\mathbb{E}_{t,x}\left[V_i\left(\tau, X_\tau^i\right)\right] - V_i\left(t, X_t^i\right) = \mathbb{E}_{t,x}\left[\int_t^\tau \mathcal{L}^{c_s^i, w_s^{i\cdot}} V_i(s, X_{s-}^i) ds\right].$$

Take the supremum on both sides over the admissible controls $(c_\cdot^i, w_\cdot^{i\cdot}) \in \mathcal{A}_{t,T}^i$. Recall that the dynamic programming principle in (3.16) implies that for any stopping time τ, we have

$$\sup_{(c^i_\cdot, w^{i\cdot})\in\mathcal{A}^i_{t,\tau}} \mathbb{E}_{t,x}\left[V_i\left(\tau, X^i_\tau\right)\right] = V_i\left(t, X^i_t\right).$$

Therefore, we arrive at:

$$0 = \sup_{(c^i_\cdot, w^{i\cdot})\in\mathcal{A}^i_{t,\tau}} \mathbb{E}_{t,x}\left[\int_t^\tau \mathcal{L}^{c^i_s, w^{i\cdot}_s} V_i(s, X^i_{s-})ds\right]. \tag{3.21}$$

We note that this step required existence of an optimal control. For small enough δ and ϵ in (3.20), we will have $\tau = t + \delta$. Therefore, (3.21) yields

$$0 = \sup_{(c^i_\cdot, w^{i\cdot})} \lim_{\delta\to 0} \frac{1}{\delta}\mathbb{E}_{t,x}\left[\int_t^{t+\delta} \mathcal{L}^{c^i_s, w^{i\cdot}_s} V_i(s, X^i_{s-})ds\right].$$

Finally, applying the Dominated Convergence Theorem gives

$$0 = \sup_{(c^i_\cdot, w^{i\cdot})} \mathcal{L}^{c_i, w_{i\cdot}} V_i(t, x),$$

which equals (3.5) after plugging in the definition of $\mathcal{L}^{c_i, w_{i\cdot}}$ from (3.19). □

Proof of Proposition 3.2. Both parts of this Proposition are proved nearly identically. For conciseness, full detail is only provided for case (i) where $\gamma_i = 1$.

(i): We first show that (3.5) has a separable solution. Next, the internal optimization problem is shown to be convex, and its objective function strictly concave. Finally, we show that the proposed solution is optimal.

Separability of the PDE: First we show the value function is separable. Plugging the ansatz $V_i(t, x) = g_i(t) + \log x$ into (3.5) and performing some simplification, we have:

$$0 = g_i'(t) + \sup_{c_i, w_{i\cdot}} \left\{(1 - c_i)\, r + \sum_{j\neq i} w_{ij}\mu_j + \frac{\eta_i\mu_i}{\phi_i} + \theta_i \bar{F}_i(c_i)\log\left(\frac{x - \eta_i x}{x}\right) \sum_{j\neq i} \theta_j \bar{F}_j(c_j)\log\left(\frac{x - \phi_j w_{ij} x}{x}\right)\right\}.$$

Observe we can cancel out all remaining x's, and obtain the following ODE for g_i:

$$0 = g_i'(t) + \frac{\eta_i \mu_i}{\phi_i} + \sup_{c_i, w_{i\cdot}} \left\{ (1-c_i)r + \sum_{j \neq i} w_{ij}\mu_j + \theta_i \bar{F}_i(c_i)\log(1-\eta_i) + \sum_{j \neq i} \theta_j \bar{F}_j(c_j)\log(1-\phi_j w_{ij}) \right\} \quad (3.22)$$

with terminal condition $g_i(T) = 0$. If $\hat{c}_i$ and $\hat{w}_{ij}$ are indeed the optimal solutions to the maximization in (3.22), g_i solves $g_i'(t) = -J_i^*$ with $g_i(T) = 0$, to which the solution is $g_i(t) = (T-t)J_i^*$ as desired.

Strict Concavity: Now we analyze the resulting optimization problem for $c_i, w_{i\cdot}$. Let $\mathcal{A}_i = \mathbb{R}_+ \times \prod_{j \neq i}[0, \phi_j^{-1})$ be the feasible set for this optimization problem. Clearly, $\mathcal{A}_i$ is a convex set. We aim to solve

$$\sup_{(c_i, w_{i\cdot}) \in \mathcal{A}_i} \quad (1-c_i)r + \sum_{j \neq i} w_{ij}\mu_j + \theta_i \bar{F}_i(c_i)\log(1-\eta_i) + \sum_{j \neq i} \theta_j \bar{F}_j(c_j)\log(1-\phi_j w_{ij}). \quad (3.23)$$

Let $h(c_i, w_{i\cdot})$ denote the function to be maximized in (3.23). It is critical to observe that h is additively separable in each of its optimization variables. Therefore, we can solve for each optimal control independently. Namely, all cross-derivatives of h equal zero, which greatly simplifies the proof of strict concavity. We begin by computing partial derivatives of h with respect to each variable, which gives

$$\begin{aligned} \frac{\partial h}{\partial c_i} &= -r - \theta_i f_i(c_i)\log(1-\eta_i) & \frac{\partial^2 h}{\partial c_i^2} &= -\theta_i f_i'(c_i)\log(1-\eta_i) \\ \frac{\partial h}{\partial w_{ij}} &= \mu_j - \phi_j \theta_j \frac{\bar{F}_j(c_j)}{1-\phi_j w_{ij}} & \frac{\partial^2 h}{\partial w_{ij}^2} &= -\phi_j^2 \theta_j \frac{\bar{F}_j(c_j)}{(1-\phi_j w_{ij})^2} \qquad \forall j \neq i. \end{aligned} \quad (3.24)$$

Observe that within $\mathcal{A}_i$, we have $(1-\phi_j w_{ij})^2 > 0$. Recall that by Assumption 3.1, the density function $f_j(\cdot)$ is fully supported on $\mathbb{R}_+$, and $f_i'(\cdot) < 0$. Therefore, it must be the case that $\bar{F}_j(c_j) > 0$ for any admissible c_j and $\partial^2_{w_{ij}, w_{ij}} h < 0$. Additionally, $\partial^2_{c_i, c_i} h < 0$ because $\eta_j > 0$.

As a result, the Hessian matrix of the objective function is negative definite in the feasible region, i.e. $\nabla^2 h \prec 0$ everywhere in $\mathcal{A}_i$. Hence h is a strictly concave function; if an optimal solution to problem (3.23) exists, it is unique (Boyd and Vandenberghe, 2004).

Optimality of Given Solution: To conclude, we must prove that (3.7) is optimal for bank i.

Note that $\frac{-r}{\theta_i \log(1-\eta_i)} > 0$. Since f_i is monotonically decreasing and positive valued on $\mathbb{R}_+$, its inverse $f_i^{-1}\left(\frac{-r}{\theta_i \log(1-\eta_i)}\right)$ is well-defined if and only if $\frac{-r}{\theta_i \log(1-\eta_i)} \leq f_i(0)$.

Since optimization problem (3.23) is convex, the first-order condition for constrained optimization is sufficient. We need only check that $y^* = (\hat{c}_i, w^*_{i\cdot}) \in \mathcal{A}_i$ satisfies

$$\nabla h(y^*)^T (y - y^*) \leq 0, \ \forall y \in \mathcal{A}_i.$$

The optimization problem for h is additively separable, so this condition is equivalent to the following.

$$\begin{aligned} \partial_{c_i} h(\hat{c}_i)(c_i - \hat{c}_i) \leq 0, \ \forall c_i \in \mathbb{R}_+, \\ \partial_{w_{ij}} h(\hat{w}_{ij})(w_{ij} - \hat{w}_{ij}) \leq 0, \ \forall w_{ij} \in \left[0, \phi_j^{-1}\right), \quad \forall j \neq i. \end{aligned} \tag{3.25}$$

Note that the partial derivative $\partial_{c_i} h$ in (3.24) is a function of only c_i. The same holds for the partials with respect to each w_{ij}. Note that these derivatives will depend on c_j, but this value is not controlled by bank i. Therefore, we will omit the dependence of these derivatives on the other optimization variables.

We begin with optimality of the proposed $\hat{c}_i$. Consider the case where $\frac{-r}{\theta_i \log(1-\eta_i)} \leq f_i(0)$, and observe that $\partial_{c_i} h(\hat{c}_i) = 0$ using (3.24). As a result, this choice of $\hat{c}_i$ satisfies the first-order condition for $\hat{c}_i$ in (3.25). Conversely, let us have $\frac{-r}{\theta_i \log(1-\eta_i)} > f_i(0)$. Since f_i is assumed to be monotone decreasing, it must be the case that $\frac{-r}{\theta_i \log(1-\eta_i)} > \max_{c \in \mathbb{R}_+} f_i(c)$. Using again (3.24), we obtain that $\partial_{c_i} h(c) < 0$ for every $c \in \mathbb{R}_+$. In particular, we will have $\partial_{c_i} h(0) < 0$, and the first-order condition (3.25) is satisfied by $\hat{c}_i = 0$. The proof of optimality for $\hat{w}_{ij}$ in (3.7) follows exactly the same steps. If it is non-zero, then the proposed value solves $\partial_{w_{ij}} h(\hat{w}_{ij}) = 0$. If not, then we know that this partial derivative is negative everywhere in the feasible region for w_{ij}. Choosing $\hat{w}_{ij} = 0$ satisfies the corresponding equation in (3.25).

Concluding, we have shown that the solution given in (3.7) satisfies (3.25). Since it lies within $\mathcal{A}_i$, it is optimal for problem (3.23). Recall that strict concavity provides uniqueness of this solution. Finally, since all banks optimize concurrently, (3.7) is obtained by plugging the optimal value c_j^* into $\hat{w}_{ij}$.

(*ii*): The proof of this result will largely mirror that of part (*i*). We first check separability of the PDE. If $V_i(t,x) = g_i(t)\frac{x^{1-\gamma_i}}{1-\gamma_i}$, then we have:

$$\partial_t V_i(t,x) = g_i'(t)\frac{x^{1-\gamma_i}}{1-\gamma_i}$$
$$\partial_x V_i(t,x) = g_i(t)\frac{x^{1-\gamma_i}}{x}$$
$$V_i(t,(1-c)x) = g_i(t)\frac{x^{1-\gamma_i}}{1-\gamma_i}(1-c)^{1-\gamma_i},\ \forall c<1.$$

Plugging these expressions into (3.5) and dividing by $x^{1-\gamma_i}$ removes any spatial variables, and we are left with the following ordinary differential equation for g_i.

$$0 = \frac{g_i'(t)}{1-\gamma_i} + g_i(t)\sup_{c_i,w_{i\cdot}}\left\{(1-c_i)r + \sum_{j\neq i} w_{ij}\mu_j + \frac{\eta_i\mu_i}{\phi_i} + \theta_i\bar{F}_i(c_i)\frac{(1-\eta_i)^{1-\gamma_i}-1}{1-\gamma_i} + \sum_{j\neq i}\theta_j\bar{F}_j(c_j)\frac{(1-\phi_j w_{ij})^{1-\gamma_i}-1}{1-\gamma_i}\right\}$$
$$g_i(T) = 1.$$

Let $\hat{c}_i$ and $\hat{w}_{ij}$ be the optimal solutions to the maximization. Then we see that g_i will solve $g_i'(t) = -(1-\gamma_i)J_i^* g_i(t)$ with $g_i(T)=1$, whose solution is $g_i(t) = \exp((1-\gamma_i)(T-t)J_i^*)$.

The optimality and uniqueness of the solution in (3.7) will be proved analogously to part (i), but by analyzing a different objective function. We are now interested in:

$$\sup_{(c_i,w_{i\cdot})\in\mathcal{A}_i}\ (1-c_i)r + \sum_{j\neq i} w_{ij}\mu_j + \theta_i\bar{F}_i(c_i)\frac{(1-\eta_i)^{1-\gamma_i}-1}{1-\gamma_i} + \sum_{j\neq i}\theta_j\bar{F}_j(c_j)\frac{(1-\phi_j w_{ij})^{1-\gamma_i}-1}{1-\gamma_i}$$

Again, this optimization problem is additively separable, which will simplify the proof of strict concavity. As before, let $h(c_i, w_{i\cdot})$ denote the function to be maximized. We compute its partial derivatives to be:

$$\frac{\partial h}{\partial c_i} = -r - \theta_i f_i(c_i)\frac{(1-\eta_i)^{1-\gamma_i}-1}{1-\gamma_i} \qquad \frac{\partial^2 h}{\partial c_i^2} = -\theta_i f_i'(c_i)\frac{(1-\eta_i)^{1-\gamma_i}-1}{1-\gamma_i}$$
$$\frac{\partial h}{\partial w_{ij}} = \mu_j - \phi_j\theta_j\bar{F}_j(c_j)(1-\phi_j w_{ij})^{-\gamma_i} \qquad \frac{\partial^2 h}{\partial w_{ij}^2} = \phi_j^2\theta_j\bar{F}_j(c_j)(-\gamma_i)(1-\phi_j w_{ij})^{-\gamma_i-1}$$

Under Assumption 3.1, we will have both $\partial^2_{c_i,c_i}h < 0$ and $\partial^2_{w_{ij},w_{ij}}h < 0$, since $w_{ij} < \phi_j^{-1}$ everywhere in $\mathcal{A}_i$. Therefore, h is strictly concave on $\mathcal{A}_i$ and the optimization problem is convex.

As a result, any optimal solution must be unique.

The remaining part of the proof mirrors that of part (i). Computing the gradient of h at the candidate solution in (3.7) and using the same argument will show that the first-order conditions in (3.25) are satisfied. Since this point is feasible, it must be optimal. □

Proof of Corollary 3.3. We proceed with a standard verification argument. We need to show that if ψ is a solution to the PDE (3.5) and it is $\mathcal{C}^{1,1}([0,T),\mathbb{R}_+)$, then it is equal to the value function. Since the proposed solutions solve the PDE and they are indeed $\mathcal{C}^{1,1}$, this will conclude.

Fix $t < T$, and choose $\{c_s^i, w_s^{i\cdot}\}_{s\in[t,T]}$ be some admissible controls. We apply Itô's formula to $\psi(s, X_s^i)$ between t and some stopping time τ^n – to be chosen optimally later. This yields, using the notation introduced in the proof of Proposition 3.1, the following:

$$\begin{aligned}\psi(\tau^n, X^i_{\tau^n}) = \psi(t, X_t^i) + \int_t^{\tau^n} \mathcal{L}^{c_s^i, w_s^{i\cdot}}\psi(s, X_s^i)ds + \int_t^{\tau^n} \left[\psi\left(s, X^i_{s-} - \eta_i X^i_{s-}\right) - \psi\left(s, X^i_{s-}\right)\right] dM_s^i \\ + \sum_{j\neq i}\int_t^{\tau^n} \left[\psi\left(s, X^i_{s-} - \phi_j w_s^{ij} X^i_{s-}\right) - \psi\left(s, X^i_{s-}\right)\right] dM_s^j.\end{aligned}$$

Recall that the compensated jump process $\left\{M_t^k\right\}_{t\geq 0}$ is a martingale. Taking the expectation conditioned on $X_t^i = x$, we obtain:

$$\begin{aligned}\mathbb{E}_{t,x}\left[\psi(\tau^n, X^i_{\tau^n})\right] = \psi(t,x) &+ \mathbb{E}_{t,x}\left[\int_t^{\tau^n} \mathcal{L}^{c_s^i, w_s^{i\cdot}}\psi(s, X_s^i)ds\right] \\ &+ \mathbb{E}_{t,x}\left[\int_t^{\tau^n} \left[\psi\left(s, X^i_{s-} - \eta_i X^i_{s-}\right) - \psi\left(s, X^i_{s-}\right)\right] dM_s^i\right] \\ &+ \sum_{j\neq i}\mathbb{E}_{t,x}\left[\int_t^{\tau^n} \left[\psi\left(s, X^i_{s-} - \phi_j w_s^{ij} X^i_{s-}\right) - \psi\left(s, X^i_{s-}\right)\right] dM_s^j\right].\end{aligned}$$

If we choose $\tau^n = \left(T - \frac{1}{n}\right) \wedge \inf\left\{s \in [t,T], X_s^i \leq \frac{1}{n} \text{ or } X_s^i \geq n\right\}$, then for every n the expectation of each stochastic integral is zero and we have:

$$\mathbb{E}_{t,x}\left[\psi(\tau^n, X^i_{\tau^n})\right] = \psi(t,x) + \mathbb{E}_{t,x}\left[\int_t^{\tau^n} \mathcal{L}^{c_s^i, w_s^{i\cdot}}\psi(s, X_s^i)ds\right].$$

Taking the limit as $n \to \infty$, we will have $\tau^n \to T$. Furthermore, since ψ satisfies the terminal condition (by assumption) and everything is bounded, an application of dominated convergence yields:

$$\mathbb{E}_{t,x}\left[U_i(X_T^i)\right] = \psi(t,x) + \mathbb{E}_{t,x}\left[\int_t^T \mathcal{L}^{c_s^i, w_s^{i\cdot}}\psi(s, X_s^i)ds\right]. \tag{3.26}$$

First, we choose the controls in (3.26) to be given by the optimal solution of Proposition 3.2. Then, we will have $\mathcal{L}^{\hat{c}_s^i, \hat{w}_s^{i\cdot}}\psi(s, X_s^i) = 0$ for all $s \in [t, \tau^n]$, and consequentially:

$$\psi(t,x) = \mathbb{E}_{t,x}\left[U_i(X_T^i)\right].$$

Note that only the terminal wealth X_T^i in the right-hand side depends on the controls $(\hat{c}_s^i, \hat{w}_s^{i\cdot})$. After taking the supremum we obtain

$$\psi(t,x) \leq \sup_{\{c_s^i, w_s^{i\cdot}\}_{s\in[t,T]}} \mathbb{E}_{t,x}\left[U_i(X_T^i)\right] = V_i(t,x). \tag{3.27}$$

Next, we fix any control $(c_s^i, w_s^{i\cdot})$. Then, in (3.26) we will have $\mathcal{L}^{c_s^i, w_s^{i\cdot}}\psi(s, X_s^i) \leq 0$, and the result is:

$$\psi(t,x) \geq \mathbb{E}_{t,x}\left[U_i(X_T^i)\right].$$

Note again that only X_T^i depends on the controls. However, since this inequality holds for any admissible control we can take the supremum over both sides to give

$$\psi(t,x) \geq \sup_{\{c_s^i, w_s^{i\cdot}\}_{s\in[t,T]}} \mathbb{E}_{t,x}\left[U_i(X_T^i)\right] = V_i(t,x). \tag{3.28}$$

Combining (3.27) and (3.28) shows that $\psi = V_i$. This implies that the optimal values to the maximization problem in the PDE for ψ are indeed the optimal controls.

Since the explicit solutions given by Proposition 3.2 are once continuously differentiable in both time and space, then they are equal to the value function. □

3.A.2 Centralized Network

Proof of Proposition 3.4. This proof is only a minor adaptation of the proof of Proposition 3.1.

First, the application of Itô's formula to the value function $V(t, X_t^1, \ldots, X_t^n)$ yields more terms, but remains simple as the jump processes are mutually independent. Namely, the generator is given by

$$\mathcal{L}^{c_\cdot,w_{\cdot\cdot}}\psi = \partial_t\psi + \sum_{i=1}^{n}\Bigg(\left[(1-c_i)r + \sum_{j\neq i} w_{ij}\mu_j + \frac{\eta_i\mu_i}{\phi_i}\right] x_i\partial_{x_i}\psi$$
$$+ \theta_i\bar{F}_i(c_i)\Big[\psi(t, x_1(1-\phi_i w_{1i}), ..., x_i(1-\eta_i), ..., x_n(1-\phi_i w_{ni})) - \psi\Big]\Bigg),$$

where ψ is evaluated at $(t, x_1, ..., x_n)$ where unspecified.

Next, to apply dominated convergence, the choice of the stopping time τ must ensure that all stopped processes $X^1_\tau, \dots X^n_\tau$ are bounded away from zero. We can therefore choose:

$$\tau = (t+\delta) \wedge \min_i \left\{\inf\left\{s \in [t,T], X^i_s \leq \epsilon \text{ or } X^i_s \geq \frac{1}{\epsilon}\right\}\right\},$$

and conclude as in the previous result. □

Proof of Proposition 3.5. The outline of this proof is similar to that of Prop. 3.2, but with greater complexity, and hence requiring additional assumptions to establish our results. We begin by discussing each of these.

First, logarithmic utility functions are needed so that (3.10) admits a separable solution. We note that if the planner sought to maximize the product of banks' utilities, it would be necessary to assume that $\gamma_i \neq 1$ for all i. This assumption is used for existence of a separable solution to (3.10).

The first condition in Assumption 3.4 concerns the shock densities f_i. In particular, (3.11) is satisfied by the family of exponential distributions ($f_i(x) = \lambda_i^{-1}e^{-\frac{x}{\lambda_i}}$, for some parameter $\lambda_i > 0$) and power distributions $\left(f_i(x) = \frac{(\alpha_i^{-1}-1)x_0^{\alpha_i^{-1}-1}}{(x+x_0)^{\alpha_i^{-1}}}$, for any $x_0 > 0$ and $\alpha_i < 1\right)$. We note that this condition is not necessary for uniqueness, but is used for establishing monotonicity of a first-order condition for optimality by bounding the second derivative with an exponentially decaying function.

Finally, the inequalities on $\Gamma(\eta_i; 1)$ will ensure that either (i): strict concavity of the objective function holds, or (ii) there exists only a single solution to the necessary first-order conditions. However, these inequalities do not rule out the possibility of a corner solution of $c_i^* = 0$ or $w_{\cdot i}^* = 0$ – as shown in (3.13). Of particular interest, the optimal decentralized and centralized allocations for c_i and $w_{\cdot i}$ will coincide whenever either $c_i^* = 0$ or $w_{\cdot i}^* = 0$ in the planner's optimum.

Separability of PDE and Maximization: Recall that the PDE for the value function derived in Proposition 3.4 is:

$$0 = \partial_t V + \sup_{c_\cdot, w_{\cdot\cdot}} \left\{ \sum_{i=1}^{n} \left(\left[(1-c_i)\, r + \sum_{j \neq i} w_{ij}\mu_j + \frac{\eta_i \mu_i}{\phi_i} \right] x_i \partial_{x_i} V \right.\right.$$
$$\left.\left. + \theta_i \bar{F}_i(c_i) \Big[V(t, x_1(1-\phi_i w_{1i}), .., x_i(1-\eta_i), .., x_n(1-\phi_i w_{ni})) - V \Big] \right) \right\} \tag{3.29}$$
$$V(T, x_1, ..., x_n) = \sum_{i=1}^{n} U_i(x_i).$$

By assumption, each bank's utility function is given by $U_i(x_i) = \log x_i$, i.e. $\gamma_i = 1$ for all i. Consider the following ansatz: $V(t, x_1, .., x_n) = g(t) + \sum_i \log x_i$. Substituting into (3.29), we obtain:

$$0 = g'(t) + \sup_{c_\cdot, w_{\cdot\cdot}} \sum_{i=1}^{n} (1-c_i)\, r + \sum_{j \neq i} w_{ij}\mu_j + \frac{\eta_i \mu_i}{\phi_i} - \theta_i \bar{F}_i(c_i) \left[\Gamma(\eta_i; 1) + \sum_{j \neq i} \Gamma(\phi_i w_{ji}; 1) \right] \tag{3.30}$$

with $g(T) = 0$. The spatial variables will cancel and we are left with an ordinary differential equation for g. We now rewrite the following sum:

$$\sum_{i=1}^{n} \sum_{j \neq i} w_{ij}\mu_j = \sum_{i=1}^{n} \sum_{j \neq i} w_{ji}\mu_i.$$

Observe that for $k, j \neq i$, we will have $w_{ji} = w_{ki}$. That is, all $j \neq i$ banks will lend the same fraction of their wealth to bank i.[8] Let this fraction be denoted by $w_{\cdot i}$. This allows us to further simplify (3.30) and obtain

$$0 = g'(t) + \sum_{i=1}^{n} \frac{\eta_i \mu_i}{\phi_i} + \sup_{c_\cdot, w_{\cdot\cdot}} \sum_{i=1}^{n} (1-c_i)\, r + (n-1) w_{\cdot i}\mu_i - \theta_i \bar{F}_i(c_i) \Big[\Gamma(\eta_i; 1) + (n-1)\Gamma(\phi_i w_{\cdot i}; 1) \Big].$$

This maximization is additively separable between each pair $(c_i, w_{\cdot i})$, indexed by i. Let $\mathcal{A}_i = \mathbb{R}_+ \times [0, \phi_i^{-1})$ denote the admissible values for $(c_i, w_{\cdot i})$. Then, the optimal allocation is found by solving:

$$\sum_{i=1}^{n} \sup_{(c_i, w_{\cdot i}) \in \mathcal{A}_i} h_i(c_i, w_{\cdot i}), \tag{3.31}$$

[8] This can be seen in two ways. First, in the decentralized setting, the amount w_{ji} depended on bank j only through their risk aversion coefficient γ_j. Since in this Proposition we have assumed that $\gamma_i = 1$ for all i, the result follows. This can also be seen by computing the first-order conditions in (3.30) for w_{ji} and w_{ki}, and noticing that they are identical.

where $h_i(c_i, w_{\cdot i}) = -rc_i + (n-1)\mu_i w_{\cdot i} - \theta_i \bar{F}_i(c_i)\Big[\Gamma(\eta_i; 1) + (n-1)\Gamma(\phi_i w_{\cdot i}; 1)\Big]$ for each i.

Reduction to Univariate Optimization: We first maximize over $w_{\cdot i}$ and then c_i given the optimal $w_{\cdot i}$. Given a value of c_i, we seek to find the optimal value of $w_{\cdot i}$. We can compute

$$\begin{aligned}\frac{\partial h_i}{\partial w_{\cdot i}}(c_i, w_{\cdot i}) &= (n-1)\mu_i - (n-1)\frac{\phi_i \theta_i \bar{F}_i(c_i)}{1-\phi_i w_{\cdot i}} \\ \frac{\partial^2 h_i}{\partial w_{\cdot i}^2}(c_i, w_{\cdot i}) &= -(n-1)\frac{\phi_i^2 \theta_i \bar{F}_i(c_i)}{(1-\phi_i w_{\cdot i})^2}.\end{aligned} \tag{3.32}$$

Notice that the second derivative in this expression is always strictly negative. Hence, given c_i, the optimization problem over $w_{\cdot i}$ is strictly concave. This implies that the first-order conditions are sufficient, and that any optimal solution is unique. Let $w_{\cdot i}^*(c_i)$ denote the optimal solution given c_i. It must satisfy the following necessary first-order condition:

$$\frac{\partial h_i}{\partial w_{\cdot i}}(c_i, w_{\cdot i}^*(c_i))(w_{\cdot i} - w_{\cdot i}^*(c_i)) \leq 0, \ \forall w_{\cdot i} \in [0, \phi_i^{-1}).$$

Using (3.32), it is easy to check that this condition is satisfied by the following:

$$w_{\cdot i}^*(c_i) = \begin{cases} \frac{1}{\phi_i}\left(1 - \frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i}\right) & \text{if } \frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i} \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.33}$$

This value is uniquely defined, and exists for any choice of c_i. We then rewrite each maximization in (3.31) as

$$\sup_{(c_i, w_{\cdot i}) \in \mathcal{A}_i} h_i(c_i, w_{\cdot i}) = \sup_{c_i \geq 0} h_i^*(c_i), \tag{3.34}$$

where $h_i^*(c_i) = h_i(c_i, w_{\cdot i}^*(c_i))$.

Existence of an Optimal Solution: We now prove existence of an optimal solution to (3.34). Observe that for large enough c_i, we will have $w_{\cdot i}^*(c_i) = \frac{1}{\phi_i}\left(1 - \frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i}\right)$. For such c_i we obtain

$$\begin{aligned} h_i^*(c_i) = &- rc_i + (n-1)\mu_i \left[\frac{1}{\phi_i}\left(1 - \frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i}\right)\right] \\ &- \theta_i \bar{F}_i(c_i)\left[\Gamma(\eta_i; 1) - (n-1)\log\left(\frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i}\right)\right]. \end{aligned}$$

As $c_i \to \infty$, we will have $\bar{F}_i(c_i) \to 0$. Since we can write

$$\bar{F}_i(c_i) \log\left(\frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i}\right) = \bar{F}_i(c_i)\left[\log\left(\frac{\phi_i \theta_i}{\mu_i}\right) + \log \bar{F}_i(c_i)\right],$$

and $x \log x \underset{x \to 0}{\to} 0$, we will have $\lim_{c_i \to \infty} h_i^*(c_i) = -\infty$.

This limit is sufficient for existence of an optimal solution to (3.34). Fix some $K < 0$. Since we have shown $h_i^*(c_i) \underset{c_i \to \infty}{\to} -\infty$, we know that $\exists C \in \mathbb{R}_+ : h_i^*(c_i) < K, \ \forall c_i > C$. By continuity of h_i^*, the set $\mathcal{B} = \{c_i \in \mathbb{R}_+ : h_i^*(c_i) \geq K\}$ is compact. We can conclude by the Extreme Value Theorem that there exists a globally optimal value of h_i^* within $\mathcal{B}$. Moreover, as long as $\mathcal{B}$ is non-empty, any point in $\mathcal{B}$ achieves higher objective value than any point in its compliment. By taking K to be a large enough negative number, we can ensure that $\mathcal{B} \neq \emptyset$.

System of Equations for Optimum: The expression (3.33) gives us the second equation in the system (3.13). For the other equation, we must analyze the first-order condition for c_i in (3.31). Taking derivatives with respect to c_i, we obtain

$$\begin{aligned} \frac{\partial h_i}{\partial c_i}(c_i, w_{\cdot i}) &= -r + \theta_i f_i(c_i)\Big[\Gamma(\eta_i; 1) + (n-1)\Gamma(\phi_i w_{\cdot i}; 1)\Big] \\ \frac{\partial^2 h_i}{\partial c_i^2}(c_i, w_{\cdot i}) &= -\theta_i f_i'(c_i)\Big[\Gamma(\eta_i; 1) + (n-1)\Gamma(\phi_i w_{\cdot i}; 1)\Big]. \end{aligned} \tag{3.35}$$

Notice that the second derivative is also negative everywhere – although this does not imply that the objective function h_i is concave. We proceed similarly as before, seeking to define an optimal value of c_i for any given $w_{\cdot i}$. Let this be denoted $c_i^*(w_{\cdot i})$. It must satisfy:

$$\frac{\partial h_i}{\partial c_i}(c_i^*(w_{\cdot i}), w_{\cdot i})(c_i - c_i^*(w_{\cdot i})) \leq 0, \ \forall c_i \in \mathbb{R}_+.$$

Using (3.35), we can see that this will be satisfied whenever

$$c_i^*(w_{\cdot i}) = \begin{cases} f_i^{-1}\left(\frac{r}{\theta_i[\Gamma(\eta_i;1)+(n-1)\Gamma(\phi_i w_{\cdot i};1)]}\right) & \text{if } f_i(0) \leq \frac{r}{\theta_i[\Gamma(\eta_i;1)+(n-1)\Gamma(\phi_i w_{\cdot i};1)]} \\ 0 & \text{otherwise.} \end{cases}$$

With (3.33), we obtain the system (3.13).

Uniqueness: It remains only to show that the optimal solution to (3.34) is unique. We return to our analysis of the univariate optimization problem in (3.34). The necessary first-order condition

for optimality of c_i^* is

$$\frac{dh_i^*}{dc_i}(c_i^*)(c_i - c_i^*) \leq 0, \ \forall c_i \in \mathbb{R}_+. \tag{3.36}$$

We proceed by showing that there exists only a single c_i^* satisfying this expression, and since existence has been proved, it must be the optimal solution. Recall that $\tilde{c}_i = F_i^{-1}\left(\left[1 - \frac{\mu_i}{\phi_i \theta_i}\right]_+\right)$, and we have $w_{\cdot i}^*(c_i) = 0$ if and only if $c_i \leq \tilde{c}_i$.

The reduced objective function $h_i^*(c_i)$, after substituting in (3.33), can be written as:

$$h_i^*(c_i) = -rc_i - \theta_i \bar{F}_i(c_i)\Gamma(\eta_i; 1) + \begin{cases} (n-1)\left[\frac{\mu_i}{\phi_i} + \theta_i \bar{F}_i(c_i)\left(\log\left(\frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i}\right) - 1\right)\right] & \text{if } c_i \geq \tilde{c}_i \\ 0 & \text{otherwise.} \end{cases}$$

Taking the derivative with respect to c_i, we obtain

$$\frac{dh_i^*}{dc_i}(c_i) = -r + \theta_i f_i(c_i)\Gamma(\eta_i; 1) - \begin{cases} \theta_i f_i(c_i)(n-1)\log\left(\frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i}\right) & \text{if } c_i \geq \tilde{c}_i \\ 0 & \text{otherwise,} \end{cases}$$

and the second derivative equals

$$\frac{d^2 h_i^*}{dc_i^2}(c_i) = \theta_i f_i'(c_i)\Gamma(\eta_i; 1) + \theta_i(n-1) \begin{cases} \frac{f_i(c_i)^2}{1 - F_i(c_i)} - f_i'(c_i)\log\left(\frac{\phi_i \theta_i \bar{F}_i(c_i)}{\mu_i}\right) & \text{if } c_i \geq \tilde{c}_i \\ 0 & \text{otherwise.} \end{cases} \tag{3.37}$$

Note that we are evaluating the right derivatives at $c_i = \tilde{c}_i$, where this function is not differentiable.

In the regime $c_i < \tilde{c}_i$, we will always have $\frac{d^2 h_i^*}{dc_i 2}(c_i) < 0$. If this were also true for $c_i \geq \tilde{c}_i$, then the objective function would be strictly concave, and uniqueness would follow. We now prove that if $\frac{d^2 h_i^*}{dc_i^2}(x) < 0$, then $h_i^*(\cdot)$ is strictly concave on $[x, \infty)$. In particular, by plugging in $x = \tilde{c}_i$ we conclude uniqueness of the optimum.

Let us compute an additional derivative of $h_i^*(\cdot)$:

$$\frac{d^3h_i^*}{dc_i^3}(c_i) = \theta_i f_i''(c_i)\Gamma(\eta_i;1)$$
$$+ \theta_i(n-1)\begin{cases} \frac{f_i(c_i)^2}{\bar{F}_i(c_i)}\left[\frac{f_i(c_i)}{\bar{F}_i(c_i)} + 3\frac{f_i'(c_i)}{f_i(c_i)}\right] - f_i''(c_i)\log\left(\frac{\phi_i\theta_i\bar{F}_i(c_i)}{\mu_i}\right) & \text{if } c_i \geq \tilde{c}_i \\ 0 & \text{otherwise.} \end{cases}$$

Observe that when we have $c_i \geq \tilde{c}_i$, a bit of algebra yields

$$\frac{d^3h_i^*}{dc_i^3}(c_i) = \frac{f_i''(c_i)}{f_i'(c_i)}\frac{d^2h_i^*}{dc_i^2}(c_i) + \frac{(n-1)\theta_i f_i(c_i)^2}{\bar{F}_i(c_i)}\left[\frac{f_i(c_i)}{\bar{F}_i(c_i)} + 3\frac{f_i'(c_i)}{f_i(c_i)} - \frac{f_i''(c_i)}{f_i'(c_i)}\right].$$

If, as assumed in this Proposition, we have $\frac{f_i(c_i)}{\bar{F}_i(c_i)} + 3\frac{f_i'(c_i)}{f_i(c_i)} - \frac{f_i''(c_i)}{f_i'(c_i)} < 0$ for all $c_i \geq 0$, then it will follow that

$$\frac{d^3h_i^*}{dc_i^3}(c_i) < \frac{f_i''(c_i)}{f_i'(c_i)}\frac{d^2h_i^*}{dc_i^2}(c_i).$$

Applying Grönwall's inequality, we see that

$$\frac{d^2h_i^*}{dc_i^2}(b) < \frac{d^2h_i^*}{dc_i^2}(a)\exp\left(\int_a^b \frac{f_i''(s)}{f_i'(s)}ds\right),$$

for any $\tilde{c}_i \leq a < b$. As a consequence, if $\frac{d^2h_i^*}{dc_i^2}(a) \leq 0$, then $\frac{d^2h_i^*}{dc_i^2}(b) < 0$ for all $b > a$.

Rewriting (3.37), we obtain:

$$\frac{d^2h_i^*}{dc_i^2}(\tilde{c}_i) = \begin{cases} \theta_i f_i'(0)\left[\Gamma(\eta_i;1) - (n-1)\log\left(\frac{\phi_i\theta_i}{\mu_i}\right)\right] + \theta_i(n-1)f_i(0)^2 & \text{if } \tilde{c}_i = 0 \\ \theta_i f_i'(\tilde{c}_i)\Gamma(\eta_i;1) + \theta_i(n-1)\frac{\phi_i\theta_i f_i(\tilde{c}_i)^2}{\mu_i} & \text{otherwise.} \end{cases}$$

For i satisfying

$$\Gamma(\eta_i;1) > \begin{cases} (n-1)\left[\log\left(\frac{\phi_i\theta_i}{\mu_i}\right) - \frac{f_i(0)^2}{f_i'(0)}\right] & \text{if } \tilde{c}_i = 0 \\ -(n-1)\frac{\phi_i\theta_i f_i(\tilde{c}_i)^2}{\mu_i f_i'(\tilde{c}_i)} & \text{otherwise.} \end{cases} \tag{3.38}$$

in the assumption (3.12), we see that $\frac{d^2h_i^*}{dc_i^2}(\tilde{c}_i) < 0$. By our application of Grönwall's inequality, we can conclude that h_i^* must be strictly concave, and hence the optimum is unique.

Now, we turn to the banks i satisfying

$$\Gamma(\eta_i;1) > \begin{cases} \frac{r}{\theta_i f_i(0)} + (n-1)\log\left(\frac{\phi_i\theta_i}{\mu_i}\right) & \text{if } \tilde{c}_i = 0 \\ \\ \frac{r}{\theta_i f_i(\tilde{c}_i)} & \text{otherwise.} \end{cases} \tag{3.39}$$

We can compute:

$$\frac{dh_i^*}{dc_i}(\tilde{c}_i) = -r + \begin{cases} \theta_i f_i(0)\left[\Gamma(\eta_i;1) - (n-1)\log\left(\frac{\phi_i\theta_i}{\mu_i}\right)\right] & \text{if } \tilde{c}_i = 0 \\ \\ \theta_i f_i(\tilde{c}_i)\Gamma(\eta_i;1) & \text{otherwise.} \end{cases}$$

By (3.39), we have $\frac{dh_i^*}{dc_i}(\tilde{c}_i) > 0$. Since $\frac{d^2h_i^*}{dc_i^2}(c_i) < 0$ for all $c_i < \tilde{c}_i$, we cannot have any points satisfying the first-order condition (3.36) in $[0, \tilde{c}_i]$. However, we do know that there must exist an optimal solution, so therefore it must lie within $(\tilde{c}_i, \infty)$. At such a point c_i^*, we must have $\frac{dh_i^*}{dc_i}(c_i^*) = 0$, and also $\frac{d^2h_i^*}{dc_i^2}(c_i^*) \leq 0$.[9] By the same conclusion using Grönwall's inequality, we must have $\frac{d^2h_i^*}{dc_i^2}(c_i) < 0$, and hence $\frac{dh_i^*}{dc_i}(c_i) < 0$ for any $c_i > c_i^*$. Hence, only this choice of c_i^* will satisfy the necessary first-order conditions, and as a result it must be unique.

Since we require all i to satisfy at least one of (3.39) or (3.38), the optimal solutions to each of the n optimization problems in (3.31) must be unique. □

Proof of Corollary 3.6. The proof of this result mirrors the proof of Corollary 3.3, and therefore we omit many details.

Fix some time $t < T$, at which we have $X_t^i = x_i$. We again choose some admissible controls $\{\dot{c_s}, \ddot{w_s}\}_{s\in[t,T]}$. We then apply Itô's formula, which only differs in yielding a few more terms. Namely, we will need to use the generator defined in Section (3.A.2), and the stochastic integrands will be slightly more complex. Next, to apply dominated convergence, our choice of the stopping time τ^n must ensure that each of the wealth processes $\left\{X_s^1\right\}_{s\geq 0} \dots \left\{X_s^n\right\}_{s\geq 0}$, is bounded at time τ^n. Therefore, we choose

$$\tau^n = \left(T - \frac{1}{n}\right) \wedge \min_i \left\{\inf\left\{s \geq t, |X_s^i - X_t^i| \geq n\right\}\right\}$$

and conclude identically. □

3.A.3 Differences in Optima

Proof of Proposition 3.7. The main idea in this proof is to first establish crude bounds of:

[9] These are the two necessary conditions for optimality of c_i^* when it lies in the interior of the feasible region.

$$\hat{c}_i \leq c_i^* \leq Kn^2,$$

for a suitable choice of K. This then allows us to improve the bounds on c_i^* itself through the relationship

$$c_i^* = f_i^{-1}\left(\frac{r}{\theta_i\left[\Gamma(\eta_i;1) - (n-1)\log\left(\frac{\phi_i\theta_i\bar{F}_i(c_i^*)}{\mu_i}\right)\right]}\right),$$

using the assumptions of a super- and sub-exponential density.

Through a direct computation with the explicit solutions in Propositions 3.2 and 3.5, we can write

$$\begin{aligned}
V(t,x_1,...,x_n) - \sum_{i=1}^n V_i(t,x_i) &= (T-t)\left[J_C^* - \sum_{i=1}^n J_i^*\right] \\
&= (T-t)\sum_{i=1}^n \Big[-r(c_i^* - \hat{c}_i) + (n-1)\mu_i(w_{\cdot i}^* - \hat{w}_{\cdot i}) \\
&\quad - \theta_i\bar{F}_i(c_i^*)\big[\Gamma(\eta_i;1) + (n-1)\Gamma(\phi_i w_{\cdot i}^*)\big] \\
&\quad + \theta_i\bar{F}_i(\hat{c}_i)\big[\Gamma(\eta_i;1) + (n-1)\Gamma(\phi_i\hat{w}_{\cdot i})\big]
\end{aligned}$$

Observe that using the definitions, we have $w_{\cdot i}^* - \hat{w}_{\cdot i} = \frac{\theta_i}{\mu_i}\left(\bar{F}_i(\hat{c}_i) - \bar{F}_i(c_i^*)\right)$. Plugging this expression in and rearranging terms, we obtain:

$$\begin{aligned}
\frac{g(t) - \sum_{i=1}^n g_i(t)}{T-t} = \sum_{i=1}^n \Bigg[& -r(c_i^* - \hat{c}_i) + \theta_i\left(\bar{F}_i(\hat{c}_i) - \bar{F}_i(c_i^*)\right)\left[(n-1) + \Gamma(\eta_i;1)\right] \\
& + \theta_i(n-1)\Big[\bar{F}_i(\hat{c}_i)\Gamma(\phi_i\hat{w}_{\cdot i};1) - \bar{F}_i(c_i^*)\Gamma(\phi_i w_{\cdot i}^*;1)\Big]\Bigg].
\end{aligned} \tag{3.40}$$

Since we know the gap in (3.40) must be positive, we can write:

$$\begin{aligned}\sum_{i=1}^{n} rc_i^* \leq & \sum_{i=1}^{n} \Bigg[r\hat{c}_i + \theta_i \left(\bar{F}_i(\hat{c}_i) - \bar{F}_i(c_i^*)\right) \left[(n-1) + \Gamma(\eta_i;1)\right] \\ & + \theta_i(n-1)\Big[\bar{F}_i(\hat{c}_i)\Gamma(\phi_i \hat{w}_{\cdot i};1) - \bar{F}_i(c_i^*)\Gamma(\phi_i w_{\cdot i}^*;1)\Big]\Bigg] \\ \leq & \sum_{i=1}^{n} \Bigg[r\hat{c}_i + \theta_i \bar{F}_i(\hat{c}_i)\big[(n-1) + \Gamma(\eta_i;1)\big] \\ & + \theta_i(n-1)\Big[\bar{F}_i(\hat{c}_i)\Gamma(\phi_i \hat{w}_{\cdot i};1)\Big]\Bigg],\end{aligned}$$

which follows by dropping the final term and since $\bar{F}_i(c_i^*) \geq 0$. A crude bound implies that

$$\begin{aligned} rc_i^* &\leq \sum_{i=1}^{n} (n-1)\Big[r\hat{c}_i + \theta_i \bar{F}_i(\hat{c}_i)\big[1 + \Gamma(\eta_i;1) + \Gamma(\phi_i \hat{w}_{\cdot i};1)\big]\Big] \\ c_i^* &\leq Kn^2, \end{aligned}$$

where $K = \max_i \left\{ \hat{c}_i + \frac{\theta_i}{r} \bar{F}_i(\hat{c}_i)\big[1 + \Gamma(\eta_i;1) + \Gamma(\phi_i \hat{w}_{\cdot i};1)\big] \right\}$ does not depend explicitly on n. Since $w_{\cdot i}^* \geq 0$, it is also easy to see that $c_i^* \geq \hat{c}_i$. Both these bounds will be useful starting points for the proof.

(i) **Upper Bound:** We first prove the upper bound for c_i^*. First, since $f_i(x) \leq \kappa_{i,U} e^{-\frac{x}{\lambda_{i,U}}}$ and both functions are decreasing, we will have $f_i^{-1}(y) \leq \lambda_{i,U} \log\left(\frac{\kappa_{i,U}}{y}\right)$, and it follows from the system of equations (3.13) that

$$c_i^* \leq \lambda_{i,U} \log\left(\frac{\theta_i \kappa_{i,U} \left[\Gamma(\eta_i;1) - (n-1)\log\left(\frac{\phi_i \theta_i}{\mu_i} \bar{F}_i(c_i^*)\right)\right]}{r} \right).$$

Now, using $f_i(x) \geq \kappa_{i,L} e^{-\frac{x}{\lambda_{i,L}}}$, we know that $\bar{F}_i(c_i^*) = \int_{c_i^*}^{\infty} f_i(u)du \geq \kappa_{i,L}\lambda_{i,L} e^{-\frac{c_i^*}{\lambda_{i,L}}}$, and write:

$$\begin{aligned} c_i^* &\leq \lambda_{i,U} \log\left(\frac{\theta_i \kappa_{i,U} \left[\Gamma(\eta_i;1) - (n-1)\log\left(\frac{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}{\mu_i}\right) + (n-1)\frac{c_i^*}{\lambda_{i,L}}\right]}{r} \right) \\ &\leq \lambda_{i,U} \log\left(\frac{\theta_i \kappa_{i,U} \left[\Gamma(\eta_i;1) - (n-1)\log\left(\frac{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}{\mu_i} \wedge 1\right) + (n-1)\frac{c_i^*}{\lambda_{i,L}}\right]}{r} \right). \end{aligned} \tag{3.41}$$

Since each of the three terms in the brackets is non-negative, we can upper bound this quantity

by:

$$c_i^* \leq \lambda_{i,U} \log\left(\frac{\theta_i \kappa_{i,U}\left[\Gamma(\eta_i;1) - \log\left(\frac{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}{\mu_i} \wedge 1\right) + \lambda_{i,L}^{-1}\right] n c_i^*}{r}\right),$$

and we define $D = \Gamma(\eta_i;1) - \log\left(\frac{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}{\mu_i} \wedge 1\right) + \lambda_{i,L}^{-1}$ for convenience. Recall that we obtained a crude upper bound of $c_i^* \leq Kn^2$, which, when plugged in, yields:

$$c_i^* \leq \lambda_{i,U} \log\left(\frac{\theta_i \kappa_{i,U} DKn^3}{r}\right).$$

This is a significantly tighter bound than Kn^2. Therefore, we plug it back into (3.41). By simplifying and bounding the term in the logarithm, we compute:

$$\begin{aligned}\frac{c_i^*}{\lambda_{i,U}} &\leq \log\left(\frac{\theta_i \kappa_{i,U}\left[\Gamma(\eta_i;1) - (n-1)\log\left(\frac{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}{\mu_i}\right) + (n-1)\frac{\lambda_{i,U}}{\lambda_{i,L}}\log\left(\frac{\theta_i \kappa_{i,U} DKn^3}{r}\right)\right]}{r}\right) \\ &\leq \log\left(\frac{\theta_i \kappa_{i,U}\left[\Gamma(\eta_i;1) + (n-1)\left[\log\left(\frac{\mu_i}{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}\left(\frac{\theta_i \kappa_{i,U} DK}{r}\right)^{\frac{\lambda_{i,U}}{\lambda_{i,L}}} \vee 1\right) + 3\frac{\lambda_{i,U}}{\lambda_{i,L}}\log(n)\right]\right]}{r}\right).\end{aligned}$$

Notice that $\Gamma(\eta_i;1) \geq 0$, $\log\left(\frac{\mu_i}{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}\left(\frac{\theta_i \kappa_{i,U} DK}{r}\right)^{\frac{\lambda_{i,U}}{\lambda_{i,L}}} \vee 1\right) \geq 0$. Therefore, we can write

$$\frac{c_i^*}{\lambda_{i,U}} \leq \log\left(\frac{\theta_i \kappa_{i,U}\left[\Gamma(\eta_i;1) + \log\left(\frac{\mu_i}{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}\left(\frac{\theta_i \kappa_{i,U} DK}{r}\right)^{\frac{\lambda_{i,U}}{\lambda_{i,L}}} \vee 1\right) + 3\frac{\lambda_{i,U}}{\lambda_{i,L}}\right](n-1)\log(n)}{r}\right),$$

and after simplification we obtain the desired bound of:

$$c_i^* \leq \lambda_{i,U} \log\left(\frac{\theta_i \kappa_{i,U} C_U}{r}\right) + \lambda_{i,U} \log\left((n-1)\log(n)\right),$$

where $C_U = \Gamma(\eta_i;1) + \log\left(\frac{\mu_i}{\phi_i \theta_i \kappa_{i,L}\lambda_{i,L}}\left(\frac{\theta_i \kappa_{i,U} DK}{r}\right)^{\frac{\lambda_{i,U}}{\lambda_{i,L}}} \vee 1\right) + 3\frac{\lambda_{i,U}}{\lambda_{i,L}}$. Observe that C_U does not depend explicitly on n, but through K it will be a function of parameters throughout the system.

Finally, it follows that $\lim_{n\to\infty} \frac{c_i^*}{\log(n)} \leq \lambda_{i,U}$.

(ii) **Lower Bound:** We proceed with the lower bound identically. With our assumption of $f_i(x) \geq \kappa_{i,L} e^{-\frac{x}{\lambda_{i,L}}}$, we know

$$c_i^* \geq \lambda_{i,L} \log\left(\frac{\theta_i \kappa_{i,L}\left[\Gamma(\eta_i;1) - (n-1)\log\left(\frac{\phi_i\theta_i}{\mu_i}\bar{F}_i(c_i^*)\right)\right]}{r}\right). \tag{3.42}$$

Moreover, since $\Gamma(\eta_i;1) \geq 0$ this term can be dropped to obtain:

$$c_i^* \geq \lambda_{i,L} \log\left(\frac{-\theta_i \kappa_{i,L}(n-1)\log\left(\frac{\phi_i\theta_i}{\mu_i}\bar{F}_i(c_i^*)\right)}{r}\right). \tag{3.43}$$

By plugging in the initial crude bound of $c_i^*\hat{c}_i$, and since $\Gamma(\phi_i\hat{w}_{\cdot i};1) = -\log\left(\frac{\phi_i\theta_i}{\mu_i}\bar{F}_i(\hat{c}_i)\right)$ by definition, we can compute a tighter lower bound for c_i^* of

$$c_i^* \geq \lambda_{i,L} \log\left(\frac{\theta_i \kappa_{i,L}(n-1)\Gamma(\phi_i\hat{w}_{\cdot i};1)}{r}\right). \tag{3.44}$$

This is precisely the lower bound in the first part of Proposition 3.7. Note that for this result, we needed only the lower bound on $f_i(\cdot)$, through which (3.42) follows.

We now continue and prove the tighter lower bound, which requires the upper bound on $f_i(\cdot)$. In particular, we assumed that $f_i(x) \leq \kappa_{i,U} e^{-\frac{x}{\lambda_{i,U}}}$, and it follows that $\bar{F}_i(c_i^*) \leq \kappa_{i,U}\lambda_{i,U} e^{-\frac{c_i^*}{\lambda_{i,U}}}$. With (3.44), we can compute an improved upper bound of:

$$\bar{F}_i(c_i^*) \leq \kappa_{i,U}\lambda_{i,U}\left(\frac{r}{\theta_i\kappa_{i,L}(n-1)\Gamma(\phi_i\hat{w}_{\cdot i})}\right)^{\frac{\lambda_{i,L}}{\lambda_{i,U}}}.$$

This upper bound on $f_i(\cdot)$ also implies that $\hat{c}_i \leq \lambda_{i,U}\log\left(\frac{\theta_i\kappa_{i,U}\Gamma(\eta_i;1)}{r}\right)$. Similarly, the assumed $f_i(x) \geq \kappa_{i,L} e^{-\frac{x}{\lambda_{i,L}}}$ will give us $\bar{F}_i(\hat{c}_i) \geq \kappa_{i,L}\lambda_{i,L} e^{-\frac{\hat{c}_i}{\lambda_{i,L}}}$. Putting the two together, we will have

$$\bar{F}_i(\hat{c}_i) \geq \kappa_{i,L}\lambda_{i,L}\left(\frac{r}{\theta_i\kappa_{i,U}\Gamma(\eta_i;1)}\right)^{\frac{\lambda_{i,U}}{\lambda_{i,L}}},$$

and it follows that

$$\bar{F}_i(c_i^*) \leq \bar{F}_i(\hat{c}_i)\frac{\kappa_{i,U}\lambda_{i,U}}{\kappa_{i,L}\lambda_{i,L}}\left(\frac{r}{\theta_i}\right)^{\frac{\lambda_{i,L}}{\lambda_{i,U}}-\frac{\lambda_{i,U}}{\lambda_{i,L}}}\frac{(\kappa_{i,U}\Gamma(\eta_i;1))^{\frac{\lambda_{i,U}}{\lambda_{i,L}}}}{(\kappa_{i,L}\Gamma(\phi_i\hat{w}_{\cdot i};1))^{\frac{\lambda_{i,L}}{\lambda_{i,U}}}}(n-1)^{-\frac{\lambda_{i,L}}{\lambda_{i,U}}}.$$

Let $C_L = \frac{\kappa_{i,U}\lambda_{i,U}}{\kappa_{i,L}\lambda_{i,L}} \left(\frac{r}{\theta_i}\right)^{\frac{\lambda_{i,L}}{\lambda_{i,U}} - \frac{\lambda_{i,U}}{\lambda_{i,L}}} \frac{(\kappa_{i,U}\Gamma(\eta_i;1))^{\frac{\lambda_{i,U}}{\lambda_{i,L}}}}{(\kappa_{i,L}\Gamma(\phi_i \hat{w}_{\cdot i};1))^{\frac{\lambda_{i,L}}{\lambda_{i,U}}}}$. Plugging this bound into (3.43), we obtain:

$$
\begin{aligned}
c_i^* &\geq \lambda_{i,L} \log\left(\frac{-\theta_i \kappa_{i,L}(n-1)\log\left(\frac{\phi_i\theta_i}{\mu_i}\bar{F}_i(\hat{c}_i) C_L (n-1)^{-\frac{\lambda_{i,L}}{\lambda_{i,U}}}\right)}{r}\right) \\
&\geq \lambda_{i,L} \log\left(\frac{-\theta_i \kappa_{i,L}(n-1)\log\left(C_L (n-1)^{-\frac{\lambda_{i,L}}{\lambda_{i,U}}}\right)}{r}\right),
\end{aligned}
$$

since $-\log\left(\frac{\phi_i\theta_i}{\mu_i}\bar{F}_i(\hat{c}_i)\right) = \Gamma(\phi_i \hat{w}_{\cdot i};1) \geq 0$, and hence this term can be dropped. Simplifying, we arrive at the desired bound of:

$$
c_i^* \geq \lambda_{i,L} \log\left(\frac{\theta_i \kappa_{i,L}\lambda_{i,L}}{r\lambda_{i,U}}\right) + \lambda_{i,L}\log\left((n-1)\Big[\log(n-1) - \frac{\lambda_{i,U}}{\lambda_{i,L}}\log(C_L)\Big]\right),
$$

from which it follows that $\lim_{n\to\infty} \frac{c_i^*}{\log(n)} \geq \lambda_{i,L}$.

Putting both (i) and (ii) together, we see that $c_i^* = \Theta\big(\log(n)\big)$. □

Proof of Proposition 3.9. Using Propositions 3.2 and 3.5, we can compute

$$
\begin{aligned}
\frac{V - \sum_{i=1}^n V_i}{T-t} = \sum_{i=1}^n \Bigg[&- r(c_i^* - \hat{c}_i) + \theta_i\left(\bar{F}_i(\hat{c}_i) - \bar{F}_i(c_i^*)\right)\big[(n-1) + \Gamma(\eta_i;1)\big] \\
&+ \theta_i(n-1)\Big[\bar{F}_i(\hat{c}_i)\Gamma(\phi_i\hat{w}_{\cdot i};1) - \bar{F}_i(c_i^*)\Gamma(\phi_i w_{\cdot i}^*;1)\Big]\Bigg],
\end{aligned}
$$

where V and V_i are evaluated at $(t, x_1, ..., x_n)$ and the difference becomes independent of wealths because of logarithmic utility. Notice that any of the terms in the sum will equal zero if $w_{\cdot i}^* = 0$ (in which case we also must also have $\hat{w}_{\cdot i} = 0$, and hence $\hat{c}_i = c_i^*$). If not, then using the results from Section 3.4 we see that

$$
\frac{-r(c_i^* - \hat{c}_i) + \theta_i\left(\bar{F}_i(\hat{c}_i) - \bar{F}_i(c_i^*)\right)\big[(n-1) + \Gamma(\eta_i;1)\big]}{n} \underset{n\to\infty}{\to} \theta_i \bar{F}_i(\hat{c}_i),
$$

since $c_i^* \asymp \log(n)$ and $\bar{F}_i(c_i^*) \to 0$. Moreover, we have seen that $(n-1)\bar{F}_i(c_i^*)\Gamma(\phi_i w_{\cdot i}^*;1) = \Theta(1)$. Since the sum is now of order $|\mathcal{M}_n|$, putting the two together yields

$$\frac{V - \sum_{i=1}^{n} V_i}{T-t} = \Theta\left(n|\mathcal{M}_n|\right).$$

In Proposition 3.2, it is easy to see that $V_i = (T-t)\Theta\left(|\mathcal{M}_n|\right)$, and therefore we obtain

$$\frac{V}{\sum_{i=1}^{n} V_i} = 1 + \Theta(1),$$

as desired. □

Proof of Corollary 3.10. This proposition is proved easily by analyzing the value functions in Propositions 3.2 and 3.5. We will use the notation of Section 3.4, where $\hat{c}_\cdot$ indicates the decentralized optimum, and $c^*_\cdot$ indicates the centralized optimum (likewise for $w_{\cdot\cdot}$).

We begin by analyzing the decentralized value function V_i. Using the explicit formula in Corollary 3.3, we write:

$$\frac{V_i}{|\mathcal{M}_n|(T-t)} = \frac{J_i^*}{|\mathcal{M}_n|} + \frac{\log x}{|\mathcal{M}_n|(T-t)},$$

and see that the second term will go to zero as $n \to \infty$. Moreover, by assumption that all banks in $\mathcal{M}_n$ are homogeneous, we will have $\hat{w}_{ij} = \hat{w}_{ik}$ for any $j, k \in M_n$. This yields:

$$J_i^* = (1-\hat{c}_i)r - \theta_i \bar{F}_i(\hat{c}_i)\Gamma(\eta_i; 1) + |\mathcal{M}_n|\left[\mu\hat{w} - \theta\bar{F}(\hat{c})\Gamma(\phi\hat{w}; 1)\right],$$

where $\hat{c}$ denotes the optimal liquidity supply held by any bank in $\mathcal{M}_n$, and $\hat{w}$ denotes the optimal investment made by any bank to those in $\mathcal{M}_n$. By using Eq (3.7) to compute $\hat{w}$, we obtain:

$$\begin{aligned} J_i^* = {} & (1-\hat{c}_i)r - \theta_i \bar{F}_i(\hat{c}_i)\Gamma(\eta_i; 1) \\ & + |\mathcal{M}_n|\left[\frac{\mu}{\phi}\left(1 - \frac{\phi\theta\bar{F}(\hat{c})}{\mu}\right) + \theta\bar{F}(\hat{c})\log\left(\frac{\phi\theta\bar{F}(\hat{c})}{\mu}\right)\right], \end{aligned}$$

and the desired limit follows.[10]

The analysis of the centralized setting is almost identical, using the value function in Proposition 3.5, we have:

$$\frac{V}{n|\mathcal{M}_n|(T-t)} = \frac{J_C^*}{n|\mathcal{M}_n|} + \frac{\sum_{i=1}^{n}\log x_i}{n|\mathcal{M}_n|(T-t)}.$$

[10] We note that this expression for J_i^* is only correct when i is not in $\mathcal{M}_n$, otherwise we would have a factor of $|\mathcal{M}_n| - 1$ in front of the term in brackets. However, in the limit this difference will vanish.

The only term of interest for large n will be J_C^*, and by homogeneity within $\mathcal{M}_n$ we can see that:

$$J_C^* = |\mathcal{M}_n|(n-1)w^*\mu + \sum_{i=1}^{n}\left((1-c_i^*)\, r - \theta_i \bar{F}_i(c_i^*)\Big[\Gamma(\eta_i;1) + (n-1)\Gamma(\phi_i w^*;1)\Big]\right),$$

where w^* denotes the optimal fractional amount invested into each bank in $\mathcal{M}_n$. Notice that only for bank in $\mathcal{M}_n$ will we have c_i^* growing with n (logarithmically). Moreover, from the analysis in Section 3.4, we also know that $(n-1)\bar{F}_i(c_i^*)\Gamma(\phi_i w_{\cdot i}^*;1)$ is of constant order. Therefore, when dividing by $n|\mathcal{M}_n|$ and taking the limit, the sum will go to zero. Only the first term will remain, and we also know that $w^* \to \phi^{-1}$ as $n \to \infty$, which concludes.

In order to show the limit for the price of anarchy, it is only necessary to sum V_i over n and divide. □

3.B Price of Anarchy: Super-/Sub-Power Distribution

In this section, we perform similar calculations to the main result of Section 3.4, but for shock size densities bounded by power law distributions. In particular, we have the following analogue of Proposition 3.7:

Proposition 3.11. *If for all x we have $f_i(x) \geq \kappa_{i,L}(\zeta_i^0 + x)^{-\frac{1}{\alpha_{i,L}}}$, for some constants $\alpha_{i,L} < 1$, $\kappa_{i,L} > 0$, and $\zeta_i^0 \geq 1$, then*

$$c_i^* \geq \left(\frac{-\kappa_{i,L}\theta_i(n-1)\log\left(\frac{\phi_i\theta_i}{\mu_i}\bar{F}_i(\hat{c}_i)\right)}{r}\right)^{\alpha_{i,L}} - \zeta_i^0.$$

If, furthermore, the density satisfies $f_i(x) \leq \kappa_{i,U}(\zeta_i^0 + x)^{-\frac{1}{\alpha_{i,U}}}$, with $\kappa_{i,U} \geq \kappa_{i,L}$ and $\alpha_{i,L} \leq \alpha_{i,U} < 1$, then:

(i) ***Upper Bound:***

$$c_i^* \leq C_U\big[(n-1)\log(n)\big]^{\alpha_{i,U}} - \zeta_i^0,$$

where C_U depends on all model parameters, but does not explicitly grow with n. As a result, $\lim_{n\to\infty} \frac{c_i^*}{\big[(n-1)\log(n)\big]^{\alpha_{i,U}}} \leq C_U$.

(ii) ***Lower Bound:***

$$c_i^* \geq \left(\frac{\kappa_{i,L}\theta_i}{r}(n-1)\left[\left(\frac{\alpha_{i,L}}{\alpha_{i,U}} - \alpha_{i,L}\right)\log(n-1) - \log(C_L)\right]\right)^{\alpha_{i,L}} - \zeta_i^0,$$

for $C_L > 0$ depending only on i. Hence, $\lim_{n\to\infty} \frac{c_i^*}{\left[(n-1)\log(n)\right]^{\alpha_{i,L}}} \geq \left(\frac{\kappa_{i,L}\theta_i}{r}\left(\frac{\alpha_{i,L}}{\alpha_{i,U}} - \alpha_{i,L}\right)\right)^{\alpha_{i,L}}$.

The proof follows an identical technique. In the special case where the shock density is indeed a power distribution, we have the following analogue of Corollary 3.8.

Corollary 3.12. *If* $f_i(x) = \frac{\left(\frac{1}{\alpha_i}-1\right)\left(\zeta_i^0\right)^{\frac{1}{\alpha_i}-1}}{\left(\zeta_i^0+x\right)^{\frac{1}{\alpha_i}}}$, *then*

$$c_i^* = \Theta\left(\left[(n-1)\log(n)\right]^{\alpha_i}\right).$$

This result can be seen by simply plugging $\alpha_{i,L} = \alpha_{i,U} = \alpha_i$ into Proposition 3.11.

This Corollary can be used to replicate the remaining analysis in Section 3.4, but as the results are qualitatively similar, we omit these calculations.

3.B.1 Proofs

Proof of Proposition 3.11. The proof of this result largely mirrors the proof of Proposition 3.7. Recall that we have shown that

$$\hat{c}_i \leq c_i^* \leq Kn^2,$$

for a suitable choice of K. By our assumptions on the density, it also follows that:

$$\left(\frac{y}{\kappa_{i,L}}\right)^{-\alpha_{i,L}} - \zeta_i^0 \leq f_i^{-1}(y) \leq \left(\frac{y}{\kappa_{i,U}}\right)^{-\alpha_{i,U}} - \zeta_i^0,$$

and

$$\frac{\kappa_{i,L}}{\frac{1}{\alpha_{i,L}}-1}\left(\zeta_i^0+x\right)^{1-\frac{1}{\alpha_{i,L}}} \leq 1 - F_i(x) \leq \frac{\kappa_{i,U}}{\frac{1}{\alpha_{i,U}}-1}\left(\zeta_i^0+x\right)^{1-\frac{1}{\alpha_{i,U}}}.$$

We can then follow the proof of Proposition 3.7 identically, but using these bounds instead. □

Chapter 4

Tradeoffs in Algorithmic Fairness

4.1 Introduction

Aspects of societal decision-making have become increasingly outsourced to algorithmic systems – including criminal risk assessment (Angwin et al., 2016), labor market organization (Chalfin et al., 2016), provision of medical care (Kleinberg et al., 2015), and more. The promise of these systems is often that they are more efficient than human decision-makers, and hence are appealing, such as for greater throughput or lower cost. However, under closer investigation, algorithmic systems have been seen to perpetuate or even amplify existing biases (Angwin et al., 2016; O'Neil, 2017; Eubanks, 2018). These concerns have brought greater attention towards the design of algorithms that exhibit fairness or other ethical qualities (Kearns and Roth, 2019; Barocas et al., 2021).

A large body of existing work defines fairness through particular statistical or mathematical quantities. These measures can be used either as constraints (Dwork et al., 2012; Hardt et al., 2016), or as a penalty imposed for deviating from equality (Berk et al., 2017). Both approaches take an egalitarian perspective, but one critique of this approach is that equality need not be fair (Cooper and Abrams, 2021). Furthermore, even if this objection is suppressed, there are many different measures of fairness (Narayanan, 2018) that can conflict with each other (Kleinberg et al., 2016). An example of this was seen within the realm of criminal justice, where journalists (Angwin et al., 2016) criticized an algorithm's unfairness with respect to one metric, while algorithm designers (Dieterich et al., 2016) demonstrated equality of another. Since the right measure of 'fairness' can be unclear, some recent research instead seeks to provide moral and ethical justification behind particular measures (Heidari et al., 2019; Hertweck et al., 2021).

Beyond the difficulty of defining and developing a 'fair' algorithm, there often exists a tension between the goals of model-builders and model-impacted individuals. For instance, algorithmic decision systems for targeted conditional cash transfer programs can exhibit comparatively greater accuracy than human-based systems, but both were found to yield inter-group inequalities (Noriega-Campero et al., 2020). On one hand, policymakers may view this result positively – they increased coverage for the needy. On the other hand, those in need may themselves desire an allocation mechanism that is fair, equal, or other such qualities – even at the expense of policymakers. From this tension has emerged a area of research aiming to understand a tradeoff between fairness and accuracy (Diana et al., 2021; Little et al., 2022; Liang et al., 2022). In particular, these studies adopt the perspective that fairness is *inherently* at odds with accuracy, which is not trivially true.[1]

Instead, this chapter identifies and formalizes a tradeoff in algorithmic design between two different ethical frameworks of distributive justice: Utilitarian (Bentham, 1996; Mill, 2008) and Rawlsian (Rawls, 2003). More precisely, we present a class of objective functions that interpolates between the preferences of a utilitarian and Rawlsian designer.

In many algorithmic settings, a model is tasked with distributing some quantity of predictive loss throughout a population. Each of these two approaches to distributive justice can be used to determine which model's allocation of loss is most 'good'. The utilitarian paradigm is often associated with accuracy (as opposed to fairness), but this need not be the case. If, for example, a utilitarian believed that each individuals' disutility is proportional to their model-induced squared error, then they would argue that the most accurate (with respect to mean squared error) model is also maximally 'good'. In doing so, moreover, this utilitarian designer weighed the needs of all individuals equally – could this not be 'fair'? It is therefore critical to emphasize that such statements about what 'fairness' is (or is not) must therefore reflect a contextual acceptance (or rejection) of particular ethical frameworks. Namely, the assertion of a 'fairness-accuracy' tradeoff requires that either: 1) utility is not tied to accuracy or 2) a utilitarian approach to the problem is unfair.

This work does not advocate for a single, universal definition of algorithmic fairness. Instead, we begin from two well-known theories of distributive justice. A utilitarian designer will define 'good' to be the sum of population utility, whereas a Rawlsian designer will measure 'good' through the outcomes of a population's least advantaged.[2] The objective functions in this chapter therefore arguably reflect a 'fairness-fairness' tradeoff – or to be precise, a 'Utilitarian good-Rawlsian good' tradeoff. In part, this tradeoff is valuable to understand because each ethical framework addresses

[1] See Cooper and Abrams (2021) for a more in-depth critique of common approaches to the 'fairness-accuracy' tradeoff.

[2] In Section 4.2 we will discuss these two theories and their implied objective functions in more detail.

a common critique of the other. A utilitarian can be indifferent towards inequality, whereas a Rawlsian's greatest concern is the most needy. Conversely, while a Rawlsian designer is unconcerned with the preferences of the majority, a utilitarian weights all individuals' preferences equally. This mixed approach allows us to partially address the shortcomings of each framework while leveraging their advantages.

Our main contributions are threefold. First, we conceptualize a class of objective functions and show that they capture a relaxation of Rawls's 'original position'. This result exhibits close ties to social welfare and risk aversion. Second, we study convergence properties of the objective functions and their minimzers. These technical results verify that we are indeed interpolating between: 1) utilitarian and Rawlsian measures of 'good', and 2) their most desirable outcomes. Finally, our experiments show the tradeoff between these two measures on several common datasets, and demonstrate how this tradeoff is influenced by model complexity. In particular, a designer's preferences (over bundles of Rawlsian and utilitarian 'good') can be used to determine their desired point along this tradeoff. In these experiments, we also study group-averaged loss, and see that an egalitarian approach may be significantly at odds with Rawlsian principles.

The rest of the chapter is organized as follows. Section 4.1.1 reviews the most relevant and recent work. Section 4.2 presents the learning problem and objective functions for utilitarian and Rawlsian designers. Section 4.3 contains our main conceptual and theoretical results, where we introduce a class of objective functions and study its properties. Section 4.4 trains various models on real-world datasets and studies several aspects of their performance. Finally, Section 4.5 concludes and presents directions for future work.

4.1.1 Relevant Literature

There are several areas of related work, each of which we present here. However, we introduce and discuss the relevant ethical theories in Section 4.2.

A significant branch of literature seeks to measure fairness through mathematical or statistical measures. These works address fairness by imposing constraints or penalties based on these measures during the in-processing stage of model building, e.g. in Dwork et al. (2012); Hardt et al. (2016); Berk et al. (2017); Corbett-Davies et al. (2017). However, there is not a universally agreed upon measure of fairness. Moreover, such formal criteria for fairness can conflict (Kleinberg et al., 2016), yield to long-term damage (Liu et al., 2018) or are subject to fundamental statistical limitations (Corbett-Davies and Goel, 2018). As a response to these challenges, recent work has grounded

particular measures of fairness in moral and ethical arguments (Heidari et al., 2019; Hertweck et al., 2021). The greatest similarity between this area and our work is a shared approach to fairness through distinct theories of ethical 'good'.

Rawls's framework has appeared in computer science literature through minimax fairness (Heidari et al., 2018; Martinez et al., 2020; Lahoti et al., 2020; Diana et al., 2021; Papadaki et al., 2022; Yang et al., 2022; Little et al., 2022). These papers focus largely on *group* minimax fairness. Instead, we study *individual* minimax fairness through a relaxation of the Rawlsian 'original position'. Two comparative advantages of our approach are: 1) avoiding any danger of fairness gerrymandering (see Kearns et al. (2018) for another solution to this issue), and 2) no requirement to be given group labels (see Hashimoto et al. (2018); Lahoti et al. (2020) for other such approaches). The most similar paper to our own is Heidari et al. (2018), where the authors use a closely related social welfare function to constrain an accuracy-maximizing optimization problem. However, we do not consider accuracy to be the fundamental objective – instead focusing on maximizing social welfare itself. Finally, we note that a different principle from Rawls's theory of justice has been studied by Liu et al. (2021), who provide techniques for imposing fair equality of opportunity on Bayesian graphical models.

Social choice theory and welfare economics have also been influenced by Rawls's principles (Sen, 1976; Hammond, 1976; D'Aspremont and Gevers, 1977). However, differing views in these areas can argue that only utilitarian designs are possible (Maskin, 1978) or rational (Harsanyi, 1975). We note that the class of social welfare functions that appear in this work and Heidari et al. (2018) are justified axiomatically by Roberts (1980). As a result, recent approaches to fairness in machine learning relying on notions of social welfare (Rambachan et al., 2020, 2021) are closely related to our work. Namely, this chapter's approach can be interpreted as a planner aiming to maximize social welfare for a particular class of functions. However, in contrast to Rambachan et al. (2021), the social welfare functions in this work do not require group-specific weights to be given a priori, and instead rely on a designer's degree of risk aversion from the Rawlsian 'original position'.

Minimax optimizations (or variants thereof) have been well-studied for their robustness qualities. For example, distributionally-robust optimization (DRO) problems can significantly improve predictive outcomes for underrepresented groups (Hashimoto et al., 2018; Sagawa et al., 2020; Li et al., 2021a). In addition, Lahoti et al. (2020) use a variant of DRO that adversarially weights observations during the learning process to improve the performance of worst-off groups, which is closely tied to the Rawlsian notion of 'good'. Most related to our work is the identical relaxation of minimax optimization known as 'Tilted Empirical Risk Minimization' (TERM), proposed by Li et al. (2021b). The authors study properties of both the loss function and its optimal solutions

under the assumption of generalized linear models. One core difference is conceptual – we focus on representing these objective functions as relaxations of the Rawlsian veil of ignorance, incorporating features of risk aversion. Technically, we rigorously prove a stronger convergence property for the optimal solutions. Finally, in our experiments we study the impact of increasing model complexity on the tradeoff between two notions of the 'good'.

Finally, a number of papers in the literature aim to characterize a 'fairness-accuracy' tradeoff (Liang et al., 2022; Little et al., 2022). We note that this tradeoff also appears in several fairness-constrained approaches (Corbett-Davies et al., 2017; Diana et al., 2021). Cooper and Abrams (2021) present a critique of these studies, questioning the assumptions that fairness and accuracy are at odds, that equality is fair, and more. We empirically study the existence of a similar tradeoff, but arguing that it instead reflects a balance between utilitarian and Rawlsian measures of 'good'. In addition, we observe how the tradeoff is affected by changes to model complexity, which to the best of our knowledge, has not been previously studied.

4.2 Modeling and Ethical Frameworks

In this section, we describe a general supervised learning setting, and two theories of distributive justice that can be used to perform model selection.

Consider a n individuals, denoted by the set $N := \{1, ..., n\}$. Let $(x_i, y_i)_{i=1...n}$ denote their observed characteristics, where $x_i \in \mathcal{X}$ are features and $y_i \in \mathcal{Y}$ is a target. In simple classification settings, we may have $\mathcal{Y} = \{0, 1\}$. A set of candidate models is defined by $f_\theta : \mathcal{X} \to \mathcal{Y}$ for each θ in parameter space Θ. Finally, we will assume that the loss function $\ell(f_\theta(x_i), y_i)$ represents the *dis*utility experienced by individual i under model f_θ.[3] For simplicity of notation, we often write this as $\ell_i(\theta)$. The function ℓ is assumed to be primitive (i.e. given a priori) and in general may be highly contextual.

Utilitarian A utilitarian designer, aligning with the political philosophy of John Stuart Mill (Mill, 2008) and Jeremy Bentham (Bentham, 1996), would seek to minimize the sum of population disutility – equivalently maximizing total utility.[4] A fundamental feature of utilitarian ethics is that it can justify harming one or more individuals if others are sufficiently compensated in doing so.

[3]Usually, a loss function is chosen only based on the task at hand and not necessarily according to this principle. However, it is possible for ℓ to reflect particular kinds of preferences – e.g. in medical diagnoses, false negatives may be more heavily penalized than false positives. The general construction of loss functions that satisfy this assumption is far beyond our scope, and must be largely driven by each model's application domain and users.

[4]Recall that we assumed the loss function $\ell(f_\theta(x_i), y_i)$ equals individual i's disutility under model f_θ. As a result, the utility-maximizing problem is equivalent to loss-minimization.

Utilitarianism also exhibits the valuable property that all individuals' needs are equally important. However, it is blind to higher-order characteristics of the distribution of utilities – i.e. large increases to its variance are justified in the name of an infinitesimal increase to its mean. As a result, it has been criticized for its indifference towards inequality (Sen, 1979). In what follows we say that a *utilitarian* designer solves:

$$\min_{\theta \in \Theta} \sum_i \ell_i(\theta), \tag{4.1}$$

whose minimizer is given by $\hat{\theta}_u$. In many common machine learning examples, the objective function may equal $n^{-1} \sum_i (y_i - f_\theta(x_i))^2$, which corresponds to a utilitarian designer with disutility equal to squared loss, i.e. $\ell_i(\theta) := (y_i - f_\theta(x_i))^2$.

Rawlsian Another possible approach comes from the thought experiment and philosophy of Rawls (2003). For completeness, we briefly state a few of his main points. First, Rawls presents the 'original position' (also referred to as the 'veil of ignorance') – wherein individuals do not know their place in society, talents, or even notions of what entails a good life. From such a position, he argues that a rational individual would desire that "inequalities are to be arranged so that they are [...] to the greatest benefit of the least advantaged" (Rawls, 2003).[5] Termed the 'difference principle', this minimax approach is desirable through its ability to address utilitarianism's indifference to inequality. However, it is an extremely strict paradigm, and is largely unconcerned with the majority of a population. In our context, a *Rawlsian* designer will aim to solve:

$$\min_{\theta \in \Theta} \max_i \ell_i(\theta), \tag{4.2}$$

whose minimzer is denoted $\hat{\theta}_r$. Practically speaking, problem (4.2) can be difficult to solve in practice.[6] In addition, a Rawlsian designer's optimal model necessarily satisfies one of two criterion: 1) if the maximization in (4.2) has a unique maximizer i, then the model has reached the fundamental limit of predictability for observation i, 2) if there are multiple maximizers, it is impossible to reduce the loss for one of these without increasing the loss of another in doing so. Effectively, this means the optimal model is agnostic to any easily-predictable observations. A Rawlsian designer therefore views

[5] A comprehensive summary of Rawls's philosophy is far beyond the aim of this work. However, we note that Rawls prioritizes two other principles before the one mentioned here. First, that all individuals are entitled to the greatest possible set of individual liberties. This principle retains a minimax flavor – if the individual with least liberties agrees to a particular organization of society, then behind the veil of ignorance, all others would necessarily agree. A second principle is that offices yielding any inequalities must be equally accessible to all (i.e. equality of opportunity). The latter principle has featured in several recent studies (Hardt et al., 2016; Liu et al., 2021).

[6] The challenge arises due to the discrete maximum between $i = 1, ..., n$, and therefore the objective is neither differentiable in i, nor on a convex domain.

outliers in a fundamentally different manner from traditional data scientists – outliers represent their assessment of good, and are not noise to be discarded. We also note that problem (4.2) is also closely tied to robust control design in engineering (Kemin and Doyle, 1998, Chapter 14) and optimization under ambiguity in economics (Gilboa and Schmeidler, 1989b) – the latter of which is reminiscent of the original position.

We remark that our presentation of Rawls's philosophy is greatly simplified. In order to justify that the solution to (4.2) is 'fair', Rawls would first require that principles be satisfied: 1) basic liberties are guaranteed and 2) offices carrying inequalities are open and equally accessible to all (see footnote 5). The latter principle on fair equality of opportunity has appeared in several recent papers (Hardt et al., 2016; Heidari et al., 2019; Liu et al., 2021). While we assume that both principles hold, systemic inequalities in society would suggest that this need not be the case. A more complete integration of Rawls' principles into the design and implementation of algorithmic systems remains a rich area for future work. It is also important to note that (4.2) reflects the result of applying Rawls's difference principle to a *relaxed* original position. Individuals must at least know their notion of good (i.e. the function ℓ), but still be unaware of all other characteristics (i.e. covariates x and target y).

Much of the literature on algorithmic fairness is interested in *group* measures of fairness. Let the given observations be partitioned into groups $G_1, ..., G_m$, which may not be mutually exclusive. Applying a minimax approach to average group loss would give:

$$\min_{\theta \in \Theta} \max_{j} \frac{1}{|G_j|} \sum_{i \in G_j} \ell_i(\theta). \tag{4.3}$$

Notice that this is an inter-group Rawlsian paradigm coupled with intra-group utilitarian approach.

It is not immediately clear which of (4.2) or (4.3) is preferred. Indeed, there is a contentious debate in the literature between *individual* and *group* fairness.[7] Individual fairness represents a limiting case of group fairness, but it can generalize poorly and be difficult to measure. Conversely, group fairness can fail to account for intra-group differences in outcomes, leading to so-called 'fairness gerrymandering' (Kearns et al., 2018).

We do not aim to resolve this debate, only to argue that the individualized approach in (4.2) is closer to reflecting Rawls's original position than (4.3). In practice it is impossible to perfectly manifest Rawls's original position – recall that the individual-driven fairness of (4.2) is still a relaxation

[7]See Dwork et al. (2012) and Sharifi-Malvajerdi et al. (2019) for examples of individual fairness, or Hardt et al. (2016) and Diana et al. (2021) for group fairness. Also see Kearns et al. (2018) on mixing both individual and group notions of fairness.

of the true veil of ignorance. An individual merely being within the sample may reflect certainty about some of their characteristics, e.g. that they are applying for a low-paying job, high-interest loan, or have been previously incarcerated. However, they remain uncertain of their characteristics within the sample – including group membership and the distribution of characteristics within each group. Now, we can instead imagine a different relaxation of the original position that is related to the group-wise approach of (4.3). Here, the group-conditional distributions of covariates X, Y must be known, while only group membership is uncertain. Individuals in this new position face strictly less uncertainty than before, and hence the veil of ignorance is more transparent. Although we focus on individual fairness, we will also empirically study the effects on groups.

4.3 Utilitarian-Rawlsian Continuum

Ultimately, the approach of both utilitarian and Rawlsian designers can have shortcomings. In the main conceptual contribution of this chapter, we define a set of objective functions that interpolates between these two seemingly conflicting paradigms.

Let $W : \mathbb{R}^n \to \mathbb{R}$ be a function used to aggregate the population's individual losses. Note that we will seek to minimize W. Consider the following properties we may desire of W.

- *Continuity:* W is continuous in each of its arguments.
- *Anonymity:* For any $v \in \mathbb{R}^n$, if σ is a permutation of N, then $W(v_1, ..., v_n) = W(v_{\sigma(1)}, ..., v_{\sigma(n)})$.
- *Strong Pareto Property:* For $v, u \in \mathbb{R}^n$, if there exists $M \subset N$ satisfying $v_j = u_j$ for all $j \in M$ and $v_i < u_i$ for all $i \in M^C$, then $W(v) < W(u)$.
- *Elimination of Indifferent Individuals:* For $v, \tilde{v}, u, \tilde{u} \in \mathbb{R}^n$, if there exists $M \subset N$ satisfying $v_j = u_j$ and $\tilde{v}_j = \tilde{u}_j$ for all $j \in M$ while $v_i = \tilde{v}_i$ and $u_i = \tilde{u}_i$ for all $i \in M^C$, then $W(v) < W(u)$ if and only if $W(\tilde{v}) < W(\tilde{u})$.
- *Pigou-Dalton Transfer Principle:* For $v, u \in \mathbb{R}^n$, if $v_k = u_k$ for any $k \in N \setminus \{i, j\}$, while both $v_i + v_j = u_i + u_j$ and $u_i < v_i < v_j < u_j$, then $W(v) \leq W(u)$.

These properties appear often in the literature on welfare economics and social choice theory (Maskin, 1978; Roberts, 1980; D'Aspremont and Gevers, 1977). They are slightly distinct here to fit the context of our problem. It is helpful to provide some intuition for the final two properties. Elimination of Indifferent Individuals asserts that when fixing a subset of coordinates, the ordering implied by W

is indifferent to the value at which those coordinates are fixed.[8] Finally, the Pigou-Dalton Principle states that all else equal, we would weakly prefer greater equality.

It is well-known that these conditions imply that W must be of the form $W(v) = \sum_i g(v_i)$ for some continuous, strictly increasing, convex function g. Informally, Elimination of Indifferent Individuals will imply that $W(v) = \sum_i g_i(v_i)$, Anonymity requires that all g_i's are identical and equal to some g, and finally the Pareto Property and Pigou-Dalton Principle yield strictly increasing and convex g, respectively. For more detail, see Maskin (1978) or Roberts (1980). For our purposes, it remains only to consider which classes of $g(\cdot)$ are permitted.

Proposition 4.1. *Let $v, u \in \mathbb{R}^n$, and $\mathbf{1}$ denote the all ones vector of appropriate dimension. If for all $\delta \in \mathbb{R}$ we have $W(v) < W(u)$ if and only if $W(v + \delta\mathbf{1}) < W(u + \delta\mathbf{1})$, then it must be that:*

$$W(v) = C \sum_i e^{\lambda v_i}$$

for some $\lambda \in \mathbb{R}$. For W to satisfy the Pareto Property and the Pigou-Dalton Principle, we must further have $C > 0$ and $\lambda > 0$.

The proof is found in Appendix 4.A. It is valuable to discuss the assumption of Proposition 4.1. It asserts that our preferences over possible loss profiles depend *only* on the differences between losses. That is, W exhibits independence of common *level*. As a consequence of this assumption, there is only a single feasible class of functions for aggregating losses, which is the sum of exponentiated individual loss.

It is possible to make a different assumption to establishes a unique form for W. Consider independence of common *scale*. That is, for any $\delta > 0$ and $u, v \in \mathbb{R}^n_+$, $W(v) < W(u) \iff W(\delta v) < W(\delta u)$. Note that in contrast with the setting of Proposition 4.1, u, v must be non-negative. In this case, we would conclude that $W = C \sum_i v_i^\alpha$, for $C > 0$ and $\alpha \geq 1$. This functional form appears in Heidari et al. (2018) as a constraint in supervised learning settings, and is axiomatically derived in the social choice literature by Roberts (1980). We note that for W to satisfy independence of both common scale and level, only a utilitarian objective function is feasible (Maskin, 1978).

We can now arrive at the following class of objective functions, which are monotone transformations of the function W appearing in Proposition 4.1.

[8] This assumption is more critical in social choice settings, wherein an individual's v_i's has a greater degree of flexibility. Here, we will use the same function ℓ for *all* losses, and hence one of M or M^C must be empty, and the assumption trivially holds.

Definition 4.1 (Utilitarian-Rawlsian Objective, $L(\theta;\lambda)$)**.** *For any $\lambda \in (0,\infty)$, we define:*

$$L(\theta;\lambda) = \frac{1}{\lambda}\log\left(\frac{1}{n}\sum_i e^{\lambda \ell_i(\theta)}\right). \tag{4.4}$$

Note this transformation normalizes the objective, which can be convenient for computational approaches. However, it is not fundamentally necessary to do so. The optimization problem associated with this objective is:

$$\min_{\theta\in\Theta} L(\theta;\lambda) = \min_{\theta\in\Theta} \frac{1}{\lambda}\log\left(\frac{1}{n}\sum_i e^{\lambda \ell_i(\theta)}\right), \tag{4.5}$$

whose optimal solution is $\hat{\theta}_\lambda$. We will see in subsequent results that as λ sweeps through the range $(0,\infty)$, we obtain a sequence of 'fair' models lying on a frontier between two competing notions of good – utilitarian and Rawlsian. This is illustrated in Figure 1, where in panel (b) the indifference curves of a designer are used to determine the best model – and hence a particular value of λ.

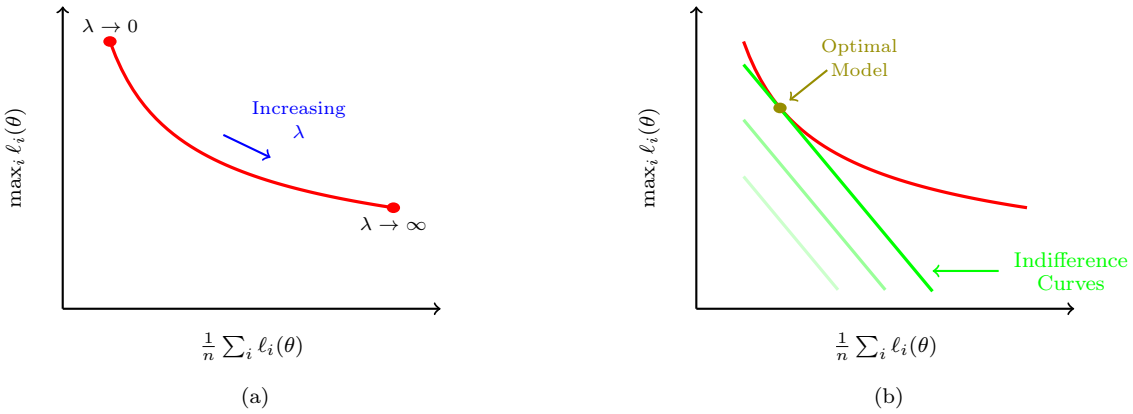


Figure 1: In panel (a), an illustration of the tradeoff between utilitarian and Rawlsian objectives. Each point on the curve represents an optimal model corresponding to some value of $\lambda \in (0,\infty)$. In (b), model selection based on societal preferences over the two types of 'good'. Greater opacity is of a particular indifference curve is associated with lower utility.

It is important to note that this framework provides one substantial difference when compared to 'fairness-as-a-constraint' approaches. Imagine that we are able to improve the expressiblity of our model class, i.e. the space Θ is strictly enlarged. Figure 2 shows that this enlargement can be used for improving either average or worst-case loss. Choosing the new optimal model should be dictated by one's relative preferences over both goods. In contrast, seeking to minimizing average loss (subject to an upper bound on maximum loss) would leverage the enlargement of Θ for only average-case performance. It is conceivable that a large improvement in worst-case loss could have been realized instead. In the experiments of Section 4.4, we will see how the frontier adjusts when

increasing model expressibility.

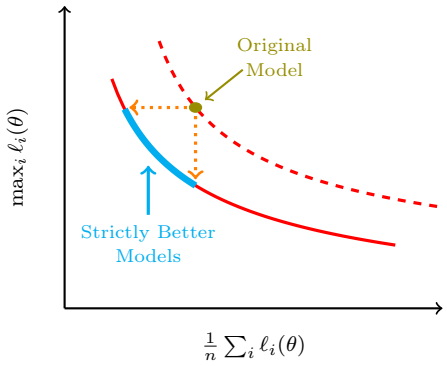


Figure 2: Illustration of how increases to model expressibility can manifest improvements to average-case and / or worst-case losses. The portion of the curve highlighted in blue represents models that are strict improvements to the original model (shown in dark green).

Main theoretical results in Section 4.3.1 show that problem (4.5) is a continuous relaxation between (4.1) and (4.2) – which is studied through both the Rawlsian original position, and convergence properties of both (4.4) and its optimal solutions. In addition, interesting connections to a regularized fairness approach and further properties of (4.5) are briefly presented in Section 4.3.2.

For completeness, we include the following analogous relaxation of the group-wise minimax approach in (4.3).

$$\min_{\theta\in\Theta} \frac{1}{\lambda} \log\left(\frac{1}{m}\sum_j e^{\lambda|G_j|^{-1}\sum_{i\in G_j}\ell_i(\theta)}\right). \tag{4.6}$$

However, as previously justified, we focus exclusively on $L(\theta;\lambda)$ and problem (4.5).

4.3.1 Characterization

This section contains our main theoretical results. We present an interpretation of $L(\theta;\lambda)$ and problem (4.5) that reflects a weakened notion of the Rawlsian veil of ignorance, and note important connections to social welfare maximization and risk aversion. Next, we study convergence properties of both $L(\theta;\lambda)$ and its minimizers $\hat{\theta}_\lambda$. Both results together verify that we are indeed representing a continuum of objective functions between utilitarian and Rawlsian designs. Finally, we conclude by briefly analyzing a simple setting – univariate linear regression.

Relaxed Veil of Ignorance and Welfare First, we show that problem (4.5) captures a natural relaxation of Rawls's original position. Consider an individual who is randomly assigned covariates X and 'true' target Y according to $X, Y \sim \text{Unif}(\{(x_1, y_1), ..., (x_n, y_n)\})$. In each state of the world θ,

she observes some random loss $\ell(f_\theta(X), Y)$. If this loss carries disutility proportional to $e^{\lambda \ell(f_\theta(X),Y)}$, then it is possible to see that:

$$\frac{1}{n}\sum_i e^{\lambda \ell_i(\theta)} = \mathbb{E}_{X,Y}\left[e^{\lambda \ell(f_\theta(X),Y)}\right].$$

In this context, the solution to problem (4.5) is equivalently minimizing expected disutility of loss for an individual with constant absolute risk aversion λ. Informally, λ captures the degree to which she dislikes uncertainty in the distribution of $\ell(f_\theta(X), Y)$. This connection is seen in the following Proposition, which is presented without proof.

Proposition 4.2. *Let $u_\lambda(\ell_i(\theta)) = -e^{\lambda \ell_i(\theta)}$ denote the utility function of individual i corresponding to model f_θ. Then:*

$$\arg\min_{\theta\in\Theta} L(\theta;\lambda) = \arg\max_{\theta\in\Theta} \mathbb{E}_{i\sim \text{Unif}[1\ldots n]}\left[u_\lambda(\ell_i(\theta))\right]. \tag{4.7}$$

Notice that the connection established in Proposition 4.2 implies that the function $L(\theta;\lambda)$ is *effectively* utilitarian – up to a monotone transformation, it is proportional to the total utility in the population. However, it is utilitarian with respect to a particular measure of utility – not loss itself. Since u_λ is a non-linear function of loss, the optimal solution to (4.7) does not coincide with the utilitarian optimum.

Convergence We now turn to the main technical results of this chapter. For limiting values of λ, we study the behavior of $L(\theta;\lambda)$ and the optimal solutions to problem (4.5).

As $\lambda \to \infty$, the sum in (4.4) is dominated by the observation with maximum loss, and hence approaches the Rawlsian minimax objective in (4.2). Conversely, as $\lambda \to 0$ the exponential is approximately linear in its argument, which leads directly to the utilitarian objective of (4.1). The following result shows that for any θ, $L(\theta;\lambda)$ indeed satisfies these properties.

Proposition 4.3. *For all $\theta \in \Theta$:*

$$\lim_{\lambda\to 0} L(\theta;\lambda) = \frac{1}{n}\sum_i \ell(f_\theta(x_i), y_i)$$

$$\lim_{\lambda\to\infty} L(\theta;\lambda) = \max_i \ell(f_\theta(x_i), y_i).$$

The proof is found in Appendix 4.A. Although simple, this result on pointwise convergence verifies that at small (resp. large) values of λ, the objective function in problem (4.5) behaves exactly like that of (4.1) (resp. (4.2)). Therefore, it is interpolating between utilitarian and Rawlsian measures

of good.

In fact, it is possible to show that $L(\theta;\lambda)$ exhibits a stronger form of convergence, which can yield convergence of its minimizers. This is formalized in the following main result.

Theorem 4.4. *Let $\hat{\theta}_\lambda$ be the optimal solution to (4.5). If $\mathcal{Y}$ is compact, the set $\{f_\theta(x), \theta \in \Theta\}$ is compact for all $x \in \mathcal{X}$, and $\ell(\cdot,\cdot)$ is continuous, then:*

$$\lim_{\lambda\to 0} \hat{\theta}_\lambda \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_i \ell(f_\theta(x_i), y_i)$$

$$\lim_{\lambda\to \infty} \hat{\theta}_\lambda \in \arg\min_{\theta\in\Theta} \max_i \ell(f_\theta(x_i), y_i).$$

In addition, if the minimizers on the right-hand side are unique (denoted $\hat{\theta}_u$ and $\hat{\theta}_r$), then:

$$\lim_{\lambda\to 0} \hat{\theta}_\lambda = \hat{\theta}_u$$

$$\lim_{\lambda\to \infty} \hat{\theta}_\lambda = \hat{\theta}_r.$$

The proof in Appendix 4.A uses the notion of Γ-convergence for a sequence of functions – which is stronger than uniform convergence. It can be leveraged to characterize the sequence of their minimizers (Braides, 2006; Maso, 2012).

Theorem 4.4 is useful for several reasons. First, it further justifies the use of $L(\theta;\lambda)$ for capturing both utilitarian and Rawlsian optimal designs. In addition, it shows that *some* minimax solution can be approximated by a sequence of minimizers to the relaxed problems. In the case where $\hat{\theta}_r$ is not unique, then we conjecture that it is possible to characterize the limit of $\hat{\theta}_\lambda$ more precisely as follows.

For $u \in \mathbb{R}^n$ let $u^{(1)}$ denote its largest entry, and $u^{(-1)}$ be the vector of remaining entries. For $u, v \in \mathbb{R}^n$, we say that $u \preceq v$ if $u^{(1)} < v^{(1)}$ or both $u^{(1)} = v^{(1)}$ and $u^{(-1)} \preceq u^{(-1)}$. This is often known as the *leximax* ordering. We expect that $\lim_{\lambda\to\infty} \hat{\theta}_\lambda = \hat{\theta}_{lex}$, where $\ell(\hat{\theta}_{lex}) \preceq \ell(\theta)$ for all θ, but to the best of our knowledge this has not yet been rigorously proven.

Example: Linear Regression We now turn to a simple setting, with $\Theta = \mathbb{R}$, $f_\theta(x) = \theta x$, and $\ell(\hat{y}, y) = (\hat{y} - y)^2$. For simplicity we also assume that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Plugging these into problem (4.5) yields the following convex and unconstrained optimization problem:

$$\min_{\theta\in\mathbb{R}} \frac{1}{\lambda} \log\left(\frac{1}{n}\sum_i e^{\lambda(\theta x_i - y_i)^2}\right).$$

The necessary (and sufficient) first-order condition can be computed as:

$$0 = \sum_i \frac{e^{\lambda(\hat{\theta}_\lambda x_i - y_i)^2}}{\sum_j e^{\lambda(\hat{\theta}_\lambda x_j - y_j)^2}} (\hat{\theta}_\lambda x_i - y_i) x_i.$$

Manipulating the above, we obtain:

$$\hat{\theta}_\lambda = \frac{\widetilde{\mathrm{Cov}}(X, Y)}{\widetilde{\mathrm{Var}}(X)},$$

which is almost exactly the usual least squares estimator. However, now the covariance and variance are computed with respect to a twisted measure $\tilde{\mathbb{P}}_\lambda$, which satisfies $\frac{d\tilde{\mathbb{P}}_\lambda}{d\mathbb{P}}(x_i) = \frac{e^{\lambda(\hat{\theta}_\lambda x_i - y_i)^2}}{n^{-1}\sum_j e^{\lambda(\hat{\theta}_\lambda x_j - y_j)^2}}$. Namely, this measure ascribes larger (resp. smaller) weights to observations whose exponentiated loss is greater (resp. less) than the average. However, it depends explicitly on $\hat{\theta}_\lambda$, and therefore the optimal solution cannot be computed in closed form.

Let us now informally consider what happens for large λ. The quantity $\sum_j e^{\lambda(\hat{\theta}_\lambda x_j - y_j)^2}$ is dominated by the observations with maximum loss, and equal measure is given to each of them. Therefore, if we let $\mathcal{I} = \arg\max_i (\hat{\theta}_\lambda x_i - y_i)^2$, then $\widetilde{\mathrm{Cov}}(X, Y) \approx \sum_{i \in \mathcal{I}} y_i x_i$ and $\widetilde{\mathrm{Var}}(X) \approx \sum_{i \in \mathcal{I}} x_i^2$. Hence, for large λ, it follows that $\hat{\theta}_\lambda \approx \frac{\mathrm{Cov}(X_{\mathcal{I}}, Y_{\mathcal{I}})}{\mathrm{Var}(X_{\mathcal{I}})}$, which is exactly the usual least squares estimator – only restricted to observations in the set $\mathcal{I}$.

We note that this setting has been more closely studied in another paper. In particular, under the assumption of generalized linear models, Li et al. (2021b) derive several interesting properties of the optimal solution. Under reasonable conditions, they show that the average loss (resp. maximum loss) is increasing (resp. decreasing) in λ at the optimal solution $\hat{\theta}_\lambda$. In addition, they prove that the empirical variance of the residuals $(\hat{\theta}_\lambda x_i - y_i)$ is non-increasing in λ, and verifies this to be the case in simulations. We might therefore expect that the finite-sample variance of $\hat{\theta}_\lambda$ is also non-decreasing in λ, although this has not been formally shown.

4.3.2 Further Properties

There are many other desirable properties of optimal solutions to learning problems, including (but not limited to) generalization performance, estimator properties, computational tractability, and optimality guarantees. In this section, we briefly touch on some of these topics and highlight connections to other areas of work – such as fairness-penalized optimization and adversarial reweighting of observations. Strengthening these results remains an active and interesting directions for future research.

Identifiability From a statistical perspective, a natural question to ask about problem (4.5) is whether or not the 'true' parameter is identifiable. That said, if the data is generated according to some $\theta^* \in \Theta$, is it possible to find θ^*? In the following result, we show that this requires a stronger condition than unbiased errors, which depends on the choice of loss function.

Proposition 4.5. *Assume that* $\exists \theta^* \in \Theta$ *for which* $Y_i|X_i \overset{i.i.d.}{\sim} f_{\theta^*}(X_i) + \epsilon_i$*, where* $\epsilon_1 ... \epsilon_n | X_i$ *are i.i.d. according to density function* f_ϵ*. Assume also that* $\ell(\hat{y}, y) = g(y - \hat{y})$ *for some differentiable and positive-valued* g*, that is strictly increasing in* $|y - \hat{y}|$*. For any* $r \in \text{Range}(g)$*, let* $g^{-1}_{(-)}(r)$ *and* $g^{-1}_{(+)}(r)$ *denote its negatively- and positively-valued inverse, respectively.*

Then, if and only if $g'\left(g^{-1}_{(+)}(r)\right) f_\epsilon\left(g^{-1}_{(+)}(r)\right) = -g'\left(g^{-1}_{(-)}(r)\right) f_\epsilon\left(g^{-1}_{(-)}(r)\right)$ *for all* $r \in \text{Range}(g)$*, then over the randomness of the sample* X, Y*, we have:*

$$\mathbb{E}\left[\nabla_\theta L(\theta^*; \lambda, X, Y)\right] = 0, \, \forall \lambda > 0.$$

The proof is found in Appendix 4.A. A special case of Proposition 4.5 occurs when both the distribution of errors f_ϵ and the primitive loss function ℓ are symmetric. In particular, given a symmetric loss function ℓ, if the distribution of errors is not symmetric, then there is no hope of obtaining a consistent estimator – the true parameter θ^* is not identifiable through the first-order conditions. However, in order to definitely prove consistency, it may be necessary to show that $L(\theta; \lambda)$ satisfies a uniform law of large numbers, which often requires compactness of Θ and that $L(\theta; \lambda)$ be bounded by a integrable function.

Regularization We now show that problem (4.5) can be used to bound an optimization problem that penalizes the objective based on its worst-case individual loss. Since for any λ and θ, $L(\theta; \lambda)$ is upper bounded (resp. lower bounded) by the maximum (resp. average) loss, there must exist some $\gamma \in (0, 1)$ for which:

$$L(\theta; \lambda) = \gamma \left(\frac{1}{n} \sum_i \ell(f_\theta(x_i), y_i) \right) + (1 - \gamma) \max_i \ell(f_\theta(x_i), y_i). \tag{4.8}$$

Fix some λ and let $\hat{\theta}_\lambda$ be the associated optimal solution to problem (4.5). We can compute its corresponding value of $\hat{\gamma}$, and majorize the following penalized optimization problem:

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_i \ell(f_\theta(x_i), y_i) + \frac{1 - \hat{\gamma}}{\hat{\gamma}} \max_i \ell(f_\theta(x_i), y_i) \leq \frac{1}{\hat{\gamma}} L(\hat{\theta}_\lambda; \lambda). \tag{4.9}$$

A similar bound can be computed in the opposite order: fix γ, minimize the γ-regularized objective (that appears in the left-hand side of (4.9)) for $\hat{\theta}_\gamma$, compute the value of $\hat{\lambda}$ that satisfies (4.8), and observe that:

$$\frac{1}{\gamma}\min_{\theta\in\Theta} L(\theta;\hat{\lambda}) \leq \frac{1}{n}\sum_i \ell\left(f_{\hat{\theta}_\gamma}(x_i), y_i\right) + \frac{1-\gamma}{\gamma}\max_i \ell\left(f_{\hat{\theta}_\gamma}(x_i), y_i\right).$$

We note that equality in the above need not hold – optimal solutions to problem (4.5) need not be minima of (4.9). In particular, $L(\theta;\lambda)$ depends on the full distribution of $\ell(\theta)$, whereas problem (4.9) is only concerned with its mean and lowest percentile. Hence, it is not always the case that problem (4.5) yields the value of θ that minimizes worst-case loss for some fixed average loss.[9] Nonetheless, a comparative benefit of problem (4.5) is that the objective function is smooth, and therefore can be solved numerically by many common algorithms.

Algorithmic Considerations From a technical perspective, $L(\theta;\lambda)$ may be preferable to the Rawlsian minimax objective because it is both differentiable and convex, which is shown in the following.

Proposition 4.6. *If $\ell(f_\theta(x), y)$ is differentiable and convex in θ for all x, y, then $L(\theta;\lambda)$ is convex.*

Proof. Since $z \to \log(\sum_i e^{\lambda z_i})$ is convex (in $z \in \mathbb{R}^n_{\geq 0}$) and non-negative, then its composition with $\ell(f_\theta(x_i), y_i)$ (which needs only be differentiable and convex) is also convex. □

As a result, we can use first-order optimization methods, which often have guaranteed convergence to a local minimum. Computing the gradient of $L(\theta;\lambda)$ gives:

$$\nabla_\theta L(\theta;\lambda) = \sum_i \frac{e^{\lambda\ell(f_\theta(x_i),y_i)}}{\sum_k e^{\lambda\ell(f_\theta(x_k),y_k)}}\nabla_\theta \ell(f_\theta(x_i), y_i), \tag{4.10}$$

where $\nabla_\theta \ell(f_\theta(x_i), y_i)$ denotes the full gradient of $\ell(f_\theta(x_i), y_i)$ with respect to θ.[10] Observe that this is simply a weighted average of the gradient at each observation i, where the weights are positively correlated with the losses. There is a relationship to adversarially re-weighted learning, for example, Lahoti et al. (2020) allows an adversarial agent to re-weight observations in order to increase a learner's weighted loss. Here, the weights are similarly related to loss, only not adversarial but pre-determined.

[9] Consider a simple example where there are two possible loss profiles (i.e. two possible values for $\ell_1(\theta), ..., \ell_n(\theta)$) given by $[0.5, 2.75, 2.75]$ and $[1, 2, 3]$. The former has smaller (resp. larger) objective value for small (resp. large) λ. However, both have the same average loss. In particular, it is possible for $[0.5, 2.75, 2.75]$ to not be the minimizer of $L(\cdot;\lambda)$.

[10] For brevity, we omit the gradient of f_θ that would appear from the chain rule.

In effect, steps along the gradient in (4.10) reflect a relaxed version of Rawls's difference principle. Originally, the principle permits inequalities only when they are to the benefit of the least advantaged. Therefore, to 'improve' over the status quo, one should aim to assist the worst-off. In (4.10), this is not necessarily the case – any harm done to the worst-off can be justified if there is *sufficient* benefit provided to others. The ability for such a setting to arise reflects a fundamental utilitarian influence. However, as λ grows, it becomes increasingly (and impossibly) difficult to justify any harm done to the worst-off.

Practically, there can be a significant computational cost associated with gradient descent. The following Proposition is from Theorem 13 in Li et al. (2021b), and slightly re-formulated here.

Proposition 4.7. *Let $\Theta \subset \mathbb{R}^d$ for some d. Assume further that for all $x, y, \theta \in \mathcal{X}, \mathcal{Y}, \Theta$, the loss function $\ell(f_\theta(x), y)$ satisfies both $||\nabla_\theta \ell(f_\theta(x), y)||_2^2 \leq C$ and*

$$C_{min} I \preccurlyeq \nabla_\theta^2 \ell(f_\theta(x), y) \preccurlyeq C_{max} I,$$

where I denotes the d-dimensional identity matrix.

Then, by running gradient descent with step size $\alpha = \frac{1}{C_{max}+2C\lambda}$, the k-th iteration $\theta^{(k)}$ satisfies:

$$L(\theta^{(k)}; \lambda) - L(\hat{\theta}_\lambda; \lambda) \leq \left(1 - \frac{C_{min}}{C_{max} + 2C\lambda}\right)^k \left(L(\theta^{(0)}; \lambda) - L(\hat{\theta}_\lambda; \lambda)\right).$$

As a direct implication, the convergence rate suffers with increases to λ. However, we might expect that problem (4.5) remains tractable for up to moderate values of λ. In addition, note that the required step size to achieve linear convergence is also decreasing in λ, which may yield further challenges. The design of efficient algorithms to solve problem (4.5) remains an open area. In our simulations, we observed that computation time was significantly reduced by using $\hat{\theta}_\lambda$ as the starting point for finding a new optimum $\hat{\theta}_{\lambda+\delta}$, for some small step $\delta > 0$.

4.4 Experiments

We now implement our methodology by solving problem (4.5) over a range of λ for several common datasets. The following can all be obtained from the UCI Machine Learning Repository (Dua and Graff, 2017).

- **COMPAS**: Arrest records from 2013 and 2014 in Broward County, Florida by (ProPublica), used in Angwin et al. (2016).

Table 1: Prediction targets and group-defining features.

Dataset	Target	Groups
COMPAS	2-Year Recidivism	Race
Bank Marketing	Subscription Decision	Marriage Status
Adult Income	Income > $50,000	Race
Credit Card Default	Payment Default	Marriage Status
Communities & Crime	Violent Crime Level	Poverty Percentage (Quartile)

Figure 3: The tradeoff between average and maximum loss at various λ for logistic regression models. Points are colored according to λ. The top panel illustrates the tradeoff for individual losses. The bottom panel shows the analogue for within-group average loss, with the identity line in gray.

- **Bank Marketing**: Part of a marketing campaign by a Portugese bank between 2008 and 2013 by Moro et al. (2014).
- **Adult Income**: Collected from the 1994 US Census, including demographic features and income.
- **Credit Card Default**: Credit card holders of a large Taiwanese bank, collected by Yeh and Lien (2009).
- **Communities & Crime**: A combination of many different features of counties within the United States, collected by Redmond and Baveja (2002). Includes sociodemographic data from the Census, survey data from law enforcement, and crime statistics collected by the FBI.

Although the objective function of (4.4) focuses on maximum individual loss, we also study average losses within groups. Table 1 shows the target variables and the group labels used in each dataset. For conciseness, all other details of our training methodology are omitted, but publicly-available here.

Average and Worst-Case Losses The tradeoff between average and worst-case loss reflects exactly the tradeoff between utilitarian and Rawlsian measures of the good. Given the sequence of optimal solutions $\{\hat{\theta}_\lambda\}_{\lambda>0}$, we compute their average and worst-case individual loss *within the training sample*. For the simple setting of logistic regression, these are shown in the top panel of Figure 3. The tradeoff is most visible for the COMPAS and Communities & Crime datasets, where as λ increases we see maximum loss reduced at the expense of average loss. However, for the other datasets worst-case performance is not significantly improved by varying λ. In fact, it appears that optimal models for the Credit Card Default dataset are indifferent to the value of λ.

We also compute average- and worst-case group loss for these datasets, where a group's loss is defined as its average – see (4.6). The bottom panel of Figure 3 plots an analogous tradeoff between average and maximum group loss for several values of λ. In the COMPAS dataset, we see that increasing λ yields a gradually more egalitarian outcome – wherein the average and maximum group losses are increased together. This suggests that equality may come at the expense of all groups.

Increasing Model Complexity We are particularly interested in studying how the curves in Figure 3 change as model complexity is increased. Intuitively, this corresponds to enlarging Θ – the set of feasible predictive models. Practically, this is associated with a greater degree of model expressibility (e.g. adding additional covariates, or training a model with greater depth). Here, we study neural networks of gradually increasing depth, and compare them to the baseline of a simple logistic regression. The main text only includes results for the COMPAS dataset, with remaining figures found in Appendix 4.B.

The top panel of Figure 4 shows that average individual loss is not significantly affected by increasing the number of layers. However, for the same value of average (individual) loss, maximum individual loss can be significantly reduced – see, for instance, the point with least maximum loss for average loss equal to 0.2. This observation suggests that when increasing model complexity, Rawlsian good may exhibit larger returns than utilitarian good. In the bottom panel of Figure 4, both average and maximum group losses greatly vary. Within this space of group losses, we often see a difference between the egalitarian (i.e. closest to the diagonal) and Rawlsian optimum. This observation suggests that equality remains at odds with both utilitarian and Rawlsian good, and in particular, that a variation of the group-skew condition from Liang et al. (2022) may hold.

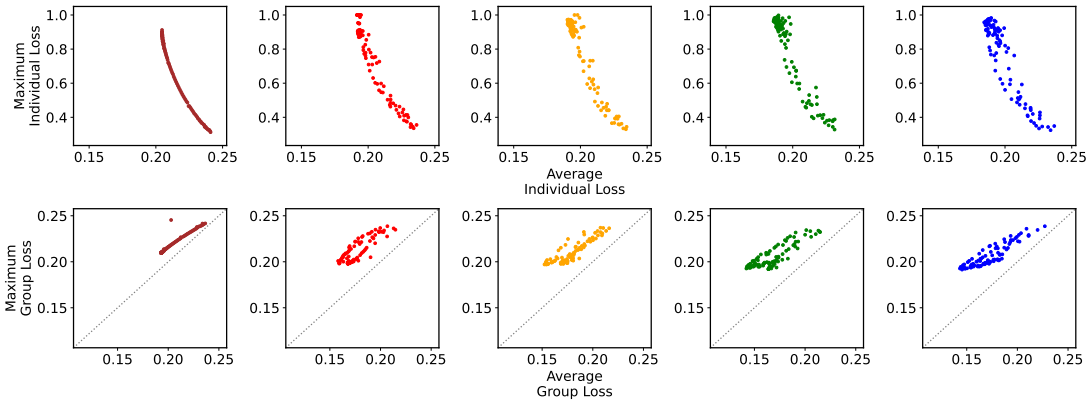


Figure 4: Result of increasing model complexity for the COMPAS dataset. From left to right: Logistic Regression, 1-Layer Neural Network,..., 4-Layer Neural Network. The top panel plots average vs maximum individual loss. The bottom panel plots average vs maximum group loss, and includes the identity line in gray for reference. In this dataset, groups were defined based on race.

4.5 Discussion and Conclusion

In this chapter, we have presented a class of objective functions for supervised learning problems that mixes aspects of both utilitarian and Rawlsian ethical frameworks. Our theoretical results are complimented by experiments on commonly-studied datasets.

Empirically, we often see a tradeoff between utilitarian and Rawlsian measures of good. From an economic perspective, this tradeoff can be interpreted as a 'production frontier' between the two goods. In this context, increasing model complexity amounts to greater production capabilities. Therefore, to determine which model along this frontier is best, it is necessary to consider a designer's preferences over fictitious bundles of 'utilitarian good' and 'Rawlsian good'. Namely, designers must determine how much utilitarian good they are willing to sacrifice for some increase in Rawlsian good. To view minimax fairness (e.g. maximum loss) as a constraint significantly reduces the richness of this question – effectively assuming that the designer's marginal rate of substitution between these two goods is infinite. Instead, we advocate for a fair-by-design perspective that incorporates broader consideration of a designer's preferences.

The objective functions in this work correspond to a relaxation of the Rawlsian original position. We have shown that this relaxation is closely tied to expected utility of a risk-averse individual facing random assignment within the population – a different veil of ignorance. In principle, it is therefore possible to choose an ideal model based on the risk appetite of model-impacted individuals. Hence, we might expect that for low-consequence decisions, risk aversion is low and less emphasis is given to the Rawlsian good. Conversely, high-consequence decision-making such as credit and criminal

justice would be strongly influenced by Rawlsian principles. Then clearly, there is unlikely to be a universally agreed upon level of fairness, which must be instead closely tied to a model's use cases.

There are several interesting and valuable directions for future work. First and foremost, we only study two particular approaches to distributive justice and their application to the in-processing stage of model building. It is also common to consider fairness during pre-processing and post-processing stages. Moreover, there are other ethical theories that can inform the development of fair models. For instance, the capability approach in Sen (1999) was developed as an alternative to utility or resource-based theories of fairness. In addition, we empirically observed a conflict between egalitarian and Rawlsian optima, which has been characterized by Liang et al. (2022) for two groups in classification settings. The further study of these theories and their potential tradeoffs remains an open area of work.

In addition, there are opportunities to further develop theory behind our so-called utilitarian-Rawlsian continuum. For example, it may be possible to develop more efficient algorithms for solving the relaxed optimization problem at large values of λ, or even for computing the optimal solutions over a wide range of λ. The statistical properties of these estimators are also of interest, as we believe that large values of λ would cause the optimal solution to have large variance (over the randomness of a set of observations). Hence it may be the case that Rawlsian good is at odds with estimation quality. Finally, characterizing the effects of increased model complexity is an extremely interesting open problem. For example, a Rawlsian designer would have no objection to including protected attributes in the training data, as they would be used only to benefit the least advantaged. It is valuable to analyze how much benefit can be gained from doing so.

Human decision-makers are uniquely endowed with the ability to entertain – but not fully accept – conflicting ethical perspectives and ideals. We hope that our work is a step towards building models that more closely reflect this ability.

Appendices

4.A Proofs

Proof of Proposition 4.1. Define $\tilde{g}(e^z) = g(z)$. Then, the level invariance assumed of g is equivalent to unit invariance of $\tilde{g}$. That is, we assumed that

$$\sum_i \tilde{g}(e^{v_i}) < \sum_i \tilde{g}(e^{u_i}) \iff \sum_i \tilde{g}(e^c e^{v_i}) < \sum_i \tilde{g}(e^c e^{u_i}).$$

For this statement to hold for all u, v, it must be the case that $\tilde{g}$ is a positively homogeneous function, and for any $t > 0$ is of the form $\tilde{g}(t) = Ct^\alpha$ for constants C and α. Conclude by seeing that $g(z) = \tilde{g}(e^z) = Ce^{\alpha z}$ as desired.

For this form of g to be increasing, it must be that $C > 0$. Finally, for convexity, it must be that $\alpha > 0$. □

Proof of Proposition 4.3. For convenience of notation, recall that we write $\ell_i(\theta) = \ell(f_\theta(x_i), y_i)$.

First, the limit for $\lambda \to 0$ is shown. Taking the limit of the expression directly yields the indeterminate form $\frac{0}{0}$, and applying L'Hôpital's Rule gives:

$$\lim_{\lambda \to 0} \sum_i \frac{e^{\lambda \ell_i(\theta)}}{\sum_j e^{\lambda \ell_j(\theta)}} \ell_i(\theta),$$

from which the desired result immediately appears.

When taking the limit $\lambda \to \infty$, another indeterminate form appears, so we again begin with:

$$\lim_{\lambda \to \infty} \sum_i \frac{e^{\lambda \ell_i(\theta)}}{\sum_j e^{\lambda \ell_j(\theta)}} \ell_i(\theta).$$

Observe that this is a weighted average, where the i-th weight is $\left(\sum_j e^{\lambda(\ell_j(\theta) - \ell_i(\theta))}\right)^{-1}$. Let us define

the set of maximizers $\mathcal{I} = \arg\max_i \ell_i(\theta)$, satisfyin $\ell_i(\theta) = \ell_*(\theta)$ for all $i \in \mathcal{I}$. For any $i \notin \mathcal{I}$, notice that $\left(\sum_j e^{\lambda(\ell_j(\theta)-\ell_i(\theta))}\right) \leq e^{-\lambda(\ell_*(\theta)-\ell_i(\theta))}$. The right-hand side converges to zero as $\lambda \to \infty$, and since all weights are lower bounded by zero, this upper bound is tight. The desired limit reduces to

$$\lim_{\lambda\to\infty} \sum_{i\in\mathcal{I}} \frac{e^{\lambda\ell_i(\theta)}}{\sum_j e^{\lambda\ell_j(\theta)}} \ell_i(\theta).$$

By a similar argument, it is possible to see that for all $i \in \mathcal{I}$, $\frac{e^{\lambda\ell_i(\theta)}}{\sum_j e^{\lambda\ell_j(\theta)}} \underset{\lambda\to\infty}{\to} \frac{1}{|\mathcal{I}|}$, and hence:

$$\lim_{\lambda\to\infty} \sum_{i\in\mathcal{I}} \frac{e^{\lambda\ell_i(\theta)}}{\sum_j e^{\lambda\ell_j(\theta)}} \ell_i(\theta) = \sum_{i\in\mathcal{I}} \frac{1}{|\mathcal{I}|} \ell_i(\theta) = \ell^*(\theta)$$

as desired. □

Proof of Theorem 4.4. Recall that we defined $\mathcal{L} = \{(\ell(f_\theta(x_1), y_1), ..., \ell(f_\theta(x_n), y_n)), \forall\theta \in \Theta\}$ as the space of all feasible loss profiles. To simplify notation, we write $\ell = (\ell_1, .., , \ell_n)$ to denote an element of $\mathcal{L}$. Since the image of Θ under $f_\theta(x)$ is assumed to be compact for every x, then continuity of $\ell(\cdot, \cdot)$ and compactness of $\mathcal{Y}$ implies compactness of $\mathcal{L}$.

First, notice that:

$$\arg\min_{\ell\in\mathcal{L}} \frac{1}{\lambda} \log\left(\frac{1}{n}\sum_i e^{\lambda\ell_i}\right) = \arg\min_{\ell\in\mathcal{L}} \frac{1}{\lambda n} \sum_i \left(e^{\lambda\ell_i} - 1\right) \tag{4.11}$$

for every λ. In particular, the minimizer of (4.11) equals $(\ell(f_{\hat\theta_\lambda}(x_1), y_1), ..., \ell(f_{\hat\theta_\lambda}(x_n), y_n))$, for $\hat\theta_\lambda$ solving problem (4.5). We define $F_\lambda(\ell) = \frac{1}{\lambda n}\sum_i \left(e^{\lambda\ell_i} - 1\right)$. Differentiating $F_\lambda(\ell)$ with respect to λ gives:

$$\frac{\partial}{\partial\lambda} F_\lambda(\ell) = \frac{\sum_i e^{\lambda\ell_i}\left(e^{-\lambda\ell_i} - (1 - \lambda\ell_i)\right)}{\lambda^2 n} \geq 0,$$

so this sequence of functions is monotone in λ. In addition, it is easy to show that for any $\ell \in \mathcal{L}$, $\lim_{\lambda\to 0} F_\lambda(\ell) = \frac{1}{n}\sum_i \ell_i$. Since this pointwise convergence holds for a monotone sequence of functions on a compact set, the convergence is uniform in $\mathcal{L}$ (Rudin, 1976, Theorem 7.13):

$$F_\lambda(\ell) \underset{\lambda\to 0}{\overset{\text{unif. in } \mathcal{L}}{\to}} \frac{1}{n}\sum_i \ell_i.$$

Furthermore, the limiting function is continuous in ℓ, so it follows that this sequence also Γ-converges in $\mathcal{L}$ (see Theorem 2.1 in Braides (2006) or Proposition 5.2 in Maso (2012)). Γ-convergence can be used to prove that the sequence of minimizers of F_λ converges to a minimizer of its Γ-limit (see

Theorem 2.10 in Braides (2006) or Corollary 7.20 in Maso (2012)). To apply these results, it is necessary to establish one additional condition on the sequence $\{F_\lambda\}$.

We say that $\{F_\lambda(\cdot)\}_{\lambda>0}$ is equi-coercive on $\mathcal{L}$ if for all $t \in \mathbb{R}$ there exists a compact set K_t for which $\{F_\lambda \leq t\} \subset K_t$ for all λ. Since $F_\lambda \geq \frac{1}{n}\sum_i \ell_i$, and the latter has compact sub-level sets on $\mathcal{L}$, then indeed $\{F_\lambda(\cdot)\}_{\lambda>0}$ is equi-coercive.

Together, equi-coercivity and Γ-convergence imply that the limit of $\{\hat{\ell}_\lambda\}_{\lambda>0}$, the sequence of minimizers to (4.11), is a minimizer to the Γ-limit of F_λ. Namely:

$$\lim_{\lambda\to 0} \operatorname*{arg\,min}_{\ell\in\mathcal{L}} F_\lambda(\ell) \in \operatorname*{arg\,min}_{\ell\in\mathcal{L}} \frac{1}{n}\sum_i \ell_i. \tag{4.12}$$

To obtain the desired result, it is only necessary to rewrite the optimization problems in terms of θ and Θ.

Of course, if the minimizer on the right-hand side is unique, then the argmax in (4.12) contains only a single value, and it must be that $\lim_{\lambda\to 0}\hat{\theta}_\lambda = \hat{\theta}_u$.

The proof for taking the limit as $\lambda \to \infty$ is nearly identical. We include its outline here.

Consider now the sequence of functions $G_\lambda(\ell) = \frac{1}{\lambda}\log\left(\sum_i e^{\lambda\ell_i}\right)$. Observe that G_λ converges pointwise to $\max_i \ell_i$. Taking a derivative with respect to λ gives:

$$\frac{\partial}{\partial\lambda}G_\lambda(\ell) = \frac{\lambda\sum_i \frac{e^{\lambda\ell_i}}{\sum_j e^{\lambda\ell_j}}\ell_i - \log\left(\sum_i e^{\lambda\ell_i}\right)}{\lambda^2} \leq \frac{\lambda\max_i \ell_i - \log(\max_i e^{\lambda\ell_i})}{\lambda^2} = 0,$$

so again this sequence is monotone. Identical arguments imply that $G_\lambda(\ell) \underset{\lambda\to\infty}{\overset{\Gamma}{\to}} \max_i \ell_i$. We can similarly use this sequence's Γ-limit to construct compact sub-level sets and prove equi-coercivity. So, we obtain

$$\lim_{\lambda\to\infty} \operatorname*{arg\,min}_{\ell\in\mathcal{L}} G_\lambda(\ell) \in \operatorname*{arg\,min}_{\ell\in\mathcal{L}} \max_i \ell_i,$$

and conclude as before. □

Proof of Proposition 4.5. First, plug in the assumption on ℓ and $\epsilon_i = y - f_{\theta^*}(x_i)$. Taking the gradient with respect to θ we have:

$$\mathbb{E}\left[\nabla_\theta L(\theta^*;\lambda,X,Y)\right] = \mathbb{E}\left[\sum_i \frac{e^{\lambda g(\epsilon_i)}}{\sum_j e^{\lambda g(\epsilon_i)}} g'(\epsilon_i)\left(-\nabla f_{\theta^*}(X_i)\right)\right].$$

By the tower property we can obtain

$$\mathbb{E}\left[\nabla_\theta L(\theta^*;\lambda,X,Y)\right] = \mathbb{E}\left[\sum_i \frac{e^{\lambda g(\epsilon_i)}}{\sum_j e^{\lambda g(\epsilon_i)}}\left(-\nabla f_{\theta^*}(x_i)\right)\mathbb{E}\left[g'(\epsilon_i)|g(\epsilon_1),...,g(\epsilon_n),X_i\right]\right].$$

Recall that ϵ_i is independent of all $g(\epsilon_j),\ j \neq i$, but $g'(\epsilon_i)$ cannot be pulled out since g^{-1} is not uniquely defined. However, since $g^{-1}(r)$ can only take two values, then $\mathbb{E}\left[g'(\epsilon_i)|g(\epsilon_i)=r,X_i\right]=0$ if and only if

$$g'\left(g_{(+)}^{-1}(r)\right)f_\epsilon\left(g_{(+)}^{-1}(r)\right) + g'\left(g_{(-)}^{-1}(r)\right)f_\epsilon\left(g_{(-)}^{-1}(r)\right) = 0,$$

for all $r \in \text{Range}(g)$, where we used the notation introduced in the Proposition. □

Proof of Proposition 4.7. Using Lemma 3 from (Li et al., 2021b), we have:

$$\begin{aligned}\nabla_\theta^2 L(\theta;\lambda) =& \sum_i \lambda e^{\lambda\left(\ell(f_\theta(x_i),y_i)-L(\theta;\lambda)\right)}\left(\nabla_\theta\ell(f_\theta(x_i),y_i)-\nabla_\theta L(\theta;\lambda)\right)\left(\nabla_\theta\ell(f_\theta(x_i),y_i)-\nabla_\theta L(\theta;\lambda)\right)^T \\ &+ e^{\lambda\left(\ell(f_\theta(x_i),y_i)-L(\theta;\lambda)\right)}\nabla_\theta^2\ell(f_\theta(x_i),y_i).\end{aligned} \tag{4.13}$$

The largest eigenvalue of this matrix can be upper bounded by Weyl's inequality as follows:

$$\begin{aligned}\lambda_{max}\left(\nabla_\theta^2 L(\theta;\lambda)\right) \leq& \lambda_{max}\left(\sum_i \lambda e^{\lambda\left(\ell(f_\theta(x_i),y_i)-L(\theta;\lambda)\right)}\left(\nabla_\theta\ell(f_\theta(x_i),y_i)-\nabla_\theta L(\theta;\lambda)\right)\left(\nabla_\theta\ell(f_\theta(x_i),y_i)-\nabla_\theta L(\theta;\lambda)\right)^T\right) \\ &+ \lambda_{max}\left(\sum_i e^{\lambda\left(\ell(f_\theta(x_i),y_i)-L(\theta;\lambda)\right)}\nabla_\theta^2\ell(f_\theta(x_i),y_i)\right).\end{aligned}$$

The second term can be upper bounded by C_{max}, since we assumed that $\nabla_\theta^2\ell(f_\theta(x),y) \preccurlyeq C_{max}I$ for all x, y and θ. The first can be controlled as follows:

$$\begin{aligned}\lambda_{max}\left(\left(\nabla_\theta\ell(f_\theta(x_i),y_i)-\nabla_\theta L(\theta;\lambda)\right)\left(\nabla_\theta\ell(f_\theta(x_i),y_i)-\nabla_\theta L(\theta;\lambda)\right)^T\right) &= ||\nabla_\theta\ell(f_\theta(x_i),y_i)-\nabla_\theta L(\theta;\lambda)||_2^2 \\ &\leq ||\nabla_\theta\ell(f_\theta(x_i),y_i)||_2^2 + ||\nabla_\theta L(\theta;\lambda)||_2^2 \\ &\leq 2C,\end{aligned}$$

since we assumed that $||\nabla_\theta\ell||_2^2 \leq C$, which itself implies that the norm of $\nabla_\theta L(\theta;\lambda)$ is bounded by the same quantity. Altogether, we arrive at:

$$\lambda_{max}\left(\nabla_\theta^2 L(\theta;\lambda)\right) \leq C_{max} + 2C\lambda.$$

By dropping the first term in (4.13), we can also obtain the lower bound of:

$$\nabla^2_\theta L(\theta;\lambda) \succcurlyeq C_{min} I.$$

Theorem 13 in (Li et al., 2021b) concludes. □

4.B Additional Figures

For conciseness, Section 4.4 on increasing model complexity only shows the results for a few datasets.

In this appendix, we include the remaining figures, along with other interesting plots.

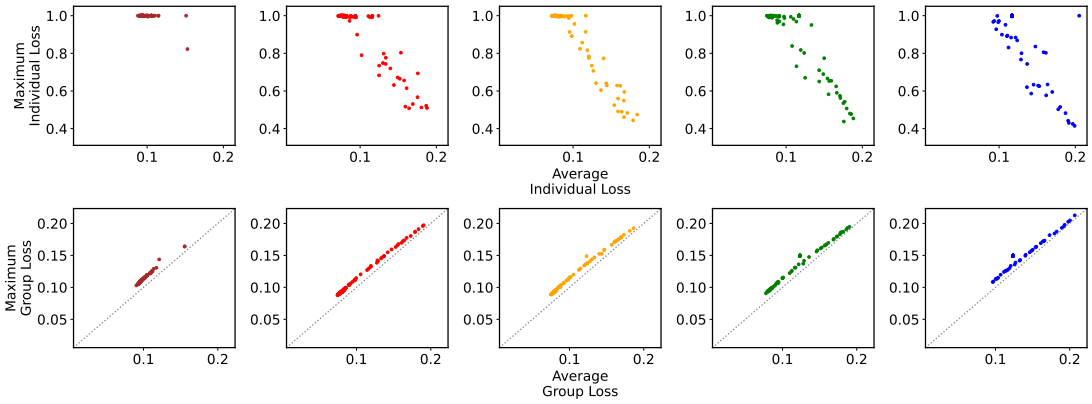


Figure 4.B.1: Result of increasing model complexity for the Bank Marketing dataset. From left to right: Logistic Regression, 1-Layer Neural Network,..., 4-Layer Neural Network. The top panel plots average vs maximum individual loss. The bottom panel plots average vs maximum group loss, and includes for reference the identity line in gray. In this dataset, groups were constructed based on an individual's marital status.

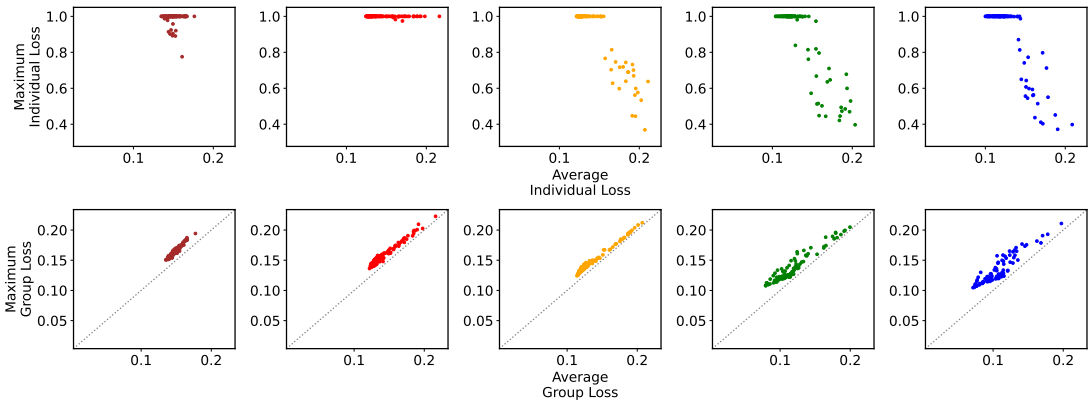


Figure 4.B.2: Result of increasing model complexity for the Adult Income dataset. From left to right: Logistic Regression, 1-Layer Neural Network,..., 4-Layer Neural Network. The top panel plots average vs maximum individual loss. The bottom panel plots average vs maximum group loss, and includes for reference the identity line in gray. In this dataset, groups were constructed based on an individual's race.

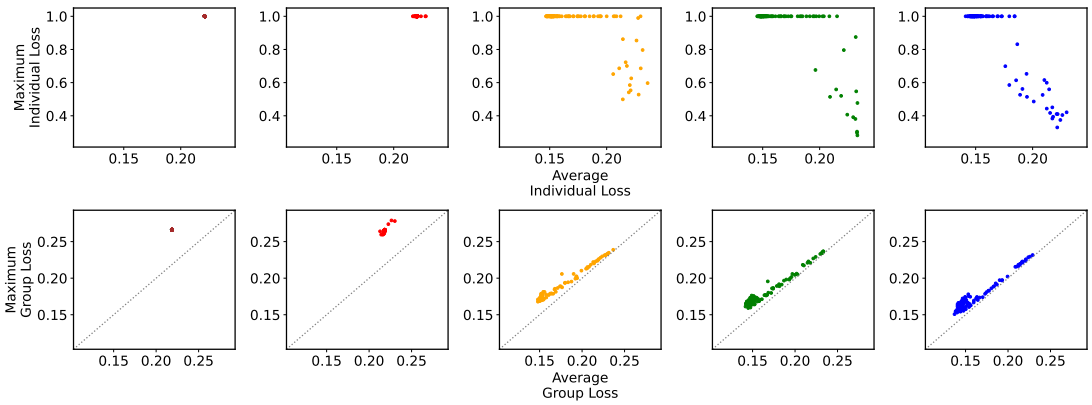


Figure 4.B.3: Result of increasing model complexity for the Credit Card Default dataset. From left to right: Logistic Regression, 1-Layer Neural Network,..., 4-Layer Neural Network. The top panel plots average vs maximum individual loss. The bottom panel plots average vs maximum group loss, and includes for reference the identity line in gray. In this dataset, groups were constructed based on an individual's marital status.

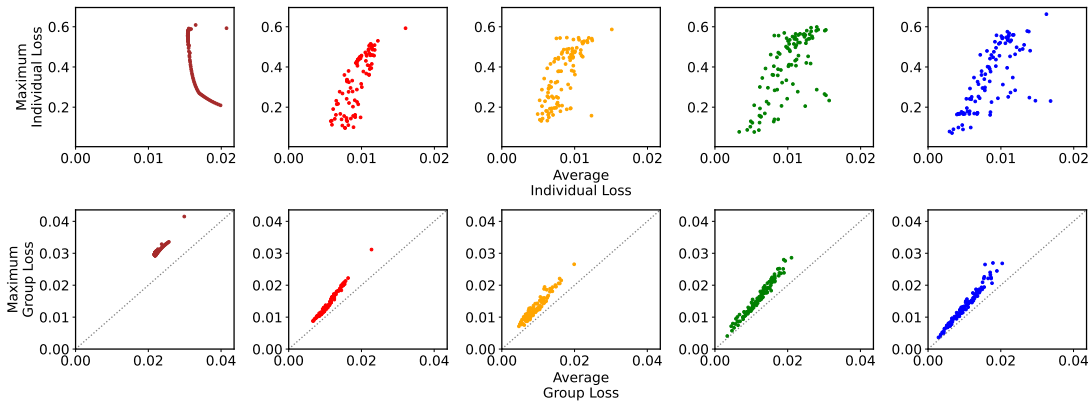


Figure 4.B.4: Result of increasing model complexity for the Communities & Crime dataset. From left to right: Logistic Regression, 1-Layer Neural Network,..., 4-Layer Neural Network. The top panel plots average vs maximum individual loss. The bottom panel plots average vs maximum group loss, and includes for reference the identity line in gray. In this dataset, groups were constructed based on quartiles of a county's poverty percentage.

Bibliography

Daron Acemoglu, Asuman E Ozdaglar, and Alireza Tahbaz-Salehi. Systemic Risk in Endogenous Financial Networks. *Columbia Business School Research Paper*, (15-17), 2015.

Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

Yacine Aït-Sahalia and T R Hurd. Portfolio Choice in Markets with Contagion. *Journal of Financial Econometrics*, 14(1):1–28, 2015.

Franklin Allen and Douglas Gale. Financial Contagion. *Journal of Political Economy*, 108(1):1–33, 2000.

Claudio Altafini. Consensus problems on networks with antagonistic interactions. *IEEE transactions on automatic control*, 58(4):935–946, 2012.

Hamed Amini and Andreea Minca. Inhomogeneous Financial Networks and Contagious Links. *Operations Research*, 64(5):1109–1120, 2016.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

Ana Babus. The Formation of Financial Networks. *The RAND Journal of Economics*, 47(2):239–272, 2016.

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

Venkatesh Bala and Sanjeev Goyal. A noncooperative model of network formation. *Econometrica*, 68(5):1181–1229, 2000.

Stefano Balietti, Lise Getoor, Daniel G Goldstein, and Duncan J Watts. Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences*, 118(52), 2021.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. 2021.

Stefano Battiston, Domenico Delli Gatti, Mauro Gallegati, Bruce Greenwald, and Joseph E. Stiglitz. Liaisons Dangereuses: Increasing Connectivity, Risk Sharing, and Systemic Risk. *Journal of Economic Dynamics and Control*, 36(8):1121–1141, 2012.

Sylvain Benoit, Jean-Edouard Colliard, Christophe Hurlin, and Christophe Pérignon. Where the Risks Lie: A Survey on Systemic Risk. *Review of Finance*, 21(1):109–152, 2017.

Sarah Bensalem, Nicolás Hernández Santibáñez, and Nabil Kazi-Tani. A continuous-time model of self-protection. *hal-02974961v2*, 2020.

Jeremy Bentham. An Introduction to the Principles of Morals and Legislation. In *The Collected Works of Jeremy Bentham*. Clarendon Press, January 1996. ISBN 978-0-19-158975-1.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A Convex Framework for Fair Regression, June 2017.

Nikita Bhalla, Adam Lechowicz, and Cameron Musco. Local edge dynamics and opinion polarization. 2021.

Bruno Biais, Thomas Mariotti, Jean-Charles Rochet, and Stéphane Villeneuve. Large Risks, Limited Liability, and Dynamic Moral Hazard. *Econometrica*, 78(1):73–118, 2010.

David Bindel, Jon Kleinberg, and Sigal Oren. How bad is forming your own opinion? *Games and Economic Behavior*, 92:248–265, 2015.

E. Biondi, C. Boldrini, A. Passarella, and M. Conti. Dynamics of opinion polarization. *arXiv preprint arXiv:2206.06134*, 2022.

Marcel Bluhm, Ester Faia, and Jan Pieter Krahnen. Endogenous Banks' Networks, Cascades and Systemic Risk. *SAFE Working Paper*, 2014.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Andrea Braides. A handbook of Γ-convergence. In M. Chipot and P. Quittner, editors, *Handbook of Differential Equations: Stationary Partial Differential Equations*, volume 3, pages 101–213. North-Holland, January 2006. doi: 10.1016/S1874-5733(06)80006-9.

Fabio Caccioli, Munik Shrestha, Cristopher Moore, and J. Doyne Farmer. Stability Analysis of Financial Contagion due to Overlapping Portfolios. *Journal of Banking and Finance*, 46(1):233–245, 2014.

Agostino Capponi and Christoph Frei. Dynamic Contracting: Accidents Lead to Nonlinear Contracts. *SIAM Journal on Financial Mathematics*, 6(1):959–983, 2015.

Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. Productivity and Selection of Human Capital with Machine Learning. *American Economic Review*, 106(5):124–127, May 2016. ISSN 0002-8282. doi: 10.1257/aer.p20161029.

Mayee Chen and Miklós Z. Rácz. An adversarial model of network disruption: Maximizing disagreement and polarization in social networks. *IEEE Transactions on Network Science and Engineering*, 9(2):728–739, 2022.

Xi Chen, Jefrey Lijffijt, and Tijl De Bie. Quantifying and Minimizing Risk of Conflict in Social Networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1197–1205, 2018.

Uthsav Chitra and Christopher Musco. Analyzing the Impact of Filter Bubbles on Social Network Polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM)*, 2020.

Fan RK Chung. *Spectral graph theory*. Number 92. American Mathematical Society, 1997.

Rodrigo Cifuentes, Gianluigi Ferrucci, and Hyun Song Shin. Liquidity Risk and Contagion. *Journal of the European Economic Association*, 3(2-3):556–566, 2005.

Ivan Conjeaud, Philipp Lorenz-Spreen, and Argyris Kalogeratos. Degroot-based opinion formation under a global steering mechanism. *arXiv preprint arXiv:2210.12274*, 2022.

Rama Cont and Peter Tankov. *Financial Modelling with Jump Processes*. Chapman and Hall/CRC, 2003.

A. Feder Cooper and Ellen Abrams. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 46–54, July 2021. doi: 10.1145/3461702.3462519.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning, August 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 797–806, New York, NY, USA, August 2017. Association for Computing Machinery. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098095.

Elisabetta Cornacchia, Neta Singer, and Emmanuel Abbe. Polarization in attraction-repulsion models. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2765–2770. IEEE, 2020.

Claude D'Aspremont and Louis Gevers. Equity and the informational basis of collective choice. *The Review of Economic Studies*, 44(2):199, June 1977. ISSN 00346527. doi: 10.2307/2297061.

Abir De, Sourangshu Bhattacharya, Parantapa Bhattacharya, Niloy Ganguly, and Soumen Chakrabarti. Learning a linear influence model from transient opinion dynamics. *Proceedings of the 2014 ACM International Conference on Information and Knowledge Management (CIKM)*, pages 401–410, 2014.

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

Nils Detering, Thilo Meyer-Brandis, Konstantinos Panagiotou, and Daniel Ritter. Managing Default Contagion in Inhomogeneous Financial Networks. *SIAM Journal on Financial Mathematics*, 10 (2):578–614, 2019.

Nils Detering, Thilo Meyer-Brandis, Konstantinos Panagiotou, and Daniel Ritter. Financial Contagion in a Stochastic Block Model. *International Journal of Theoretical and Applied Finance*, 23 (08), 2020.

Nils Detering, Thilo Meyer-Brandis, Konstantinos Panagiotou, and Daniel Ritter. An Integrated Model for Fire Sales and Default Contagion. *Mathematics and Financial Economics*, 15(1):59–101, 2021.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, March 2021.

William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpointe Inc.*, July 2016.

Luca Donetti, Franco Neri, and Miguel A. Mũoz. Optimal network topologies: Expanders, cages, Ramanujan graphs, entangled networks and all that. *Journal of Statistical Mechanics: Theory and Experiment*, (8), 2006. ISSN 17425468. doi: 10.1088/1742-5468/2006/08/P08007.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, January 2012. Association for Computing Machinery. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255.

Ronen Eldan, Miklós Z. Rácz, and Tselil Schramm. Braess's paradox for the spectral gap in random graphs and delocalization of eigenvectors. *Random Structures and Algorithms*, 50(4):584–611, 2017.

Matthew Elliott, Benjamin Golub, and Matthew O Jackson. Financial Networks and Contagion. *American Economic Review*, 104(10):3115–53, 2014.

Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group, January 2018. ISBN 978-1-4668-8596-7.

Maryam Farboodi. Intermediation and Voluntary Exposure to Counterparty Risk. *NBER Working Paper*, (w29467), 2021.

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

Prasanna Gai and Sujit Kapadia. Contagion in Financial Networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 466(2120):2401–2423, 2010.

Prasanna Gai, Andrew Haldane, and Sujit Kapadia. Complexity, Concentration and Contagion. *Journal of Monetary Economics*, 58(5):453–470, 2011.

Jason Gaitonde, Jon Kleinberg, and Eva Tardos. Adversarial Perturbations of Opinion Dynamics in Networks. In *Proceedings of the 21st ACM Conference on Economics and Computation (EC)*, pages 471–472, 2020.

Jason Gaitonde, Jon Kleinberg, and Éva Tardos. Polarization in geometric opinion dynamics. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 499–519, 2021.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 81–90, 2017.

Itzhak Gilboa and David Schmeidler. Maxmin Expected Utility with a Non-Unique Prior. *Journal of Mathematical Economics*, 18(2), 1989a.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics*, 18(2), 1989b.

Aristides Gionis, Evimaria Terzi, and Panayiotis Tsaparas. Opinion maximization in social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 387–395, 2013.

Benjamin Golub and Matthew O Jackson. How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338, 2012.

Shahrzad Haddadan, Cristina Menghini, Matteo Riondato, and Eli Upfal. Repbublik: Reducing polarized bubble radius with link insertions. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 139–147, 2021.

A A Hagberg, D A Schult, and P J Swart. Exploring network structure, dynamics, and function using NetworkX. *7th Python in Science Conference*, 2008.

Peter J. Hammond. Equity, arrow's conditions, and rawls' difference principle. *Econometrica*, 44 (4):793, July 1976. ISSN 00129682. doi: 10.2307/1913445.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

John C. Harsanyi. Nonlinear social welfare functions. *Theory and Decision*, 6(3):311–332, August 1975. ISSN 0040-5833, 1573-7187. doi: 10.1007/BF00136200.

Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 2018 International Conference on Machine Learning*, 2018.

Jan Hązła, Yan Jin, Elchanan Mossel, and Govind Ramnarayan. A geometric model of opinion polarization. *arXiv preprint arXiv:1910.05274*, 2019.

Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.

Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 181–190, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287584.

Nicolás Hernández Santibáñez, Dylan Possamaï, and Chao Zhou. Bank Monitoring Incentives under Moral Hazard and Adverse Selection. *Journal of Optimization Theory and Applications*, 184(3): 988–1035, 2020.

Corinna Hertweck, Christoph Heitz, and Michele Loi. On the moral justification of statistical parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 747–757, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445936.

Christopher Hoffman, Matthew Kahle, and Elliot Paquette. Spectral gaps of random graphs and applications. *International Mathematics Research Notices*, 2021(11):8353–8404, 2021.

Roger A Horn and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

Matthew O Jackson and Agathe Pernoud. Systemic Risk in Financial Networks: A Survey. *Annual Review of Economics*, 13:171–202, 2021.

Matthew O Jackson and Asher Wolinsky. A Strategic Model of Social and Economic Networks. *Journal of Economic Theory*, 71:44–74, 1996.

Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, October 2019. ISBN 978-0-19-094822-1.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2564–2572. PMLR, July 2018.

Zhou Kemin and John Comstock Doyle. *Essentials of Robust Control*, volume 104. Prentice Hall, 1998.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction Policy Problems. *American Economic Review*, 105(5):491–495, May 2015. ISSN 0002-8282. doi: 10.1257/aer.p20151023.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores, November 2016.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. Fairness without demographics through adversarially reweighted learning. In *Proceedings of the 2020 Conference on Neural Information Processing Systems*, 2020.

PA W Lewis and Gerald S Shedler. Simulation of Nonhomogeneous Poisson Processes by Thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

Mike Li, Hongseok Namkoong, and Shangzhou Xia. Evaluating model performance under worst-case subpopulations. In *Proceedings of the 2021 Conference on Neural Information Processing Systems*, 2021a.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization, March 2021b.

Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: Fairness versus accuracy. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, pages 58–59, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9150-4. doi: 10.1145/3490486.3538237.

Camille Olivia Little, Michael Weylandt, and Genevera I. Allen. To the fairness frontier and beyond: Identifying, quantifying, and optimizing the fairness-accuracy pareto frontier, May 2022.

David Liu, Zohair Shafi, William Fleisher, Tina Eliassi-Rad, and Scott Alfeld. RAWLSNET: Altering bayesian networks to encode rawlsian fair equality of opportunity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 745–755, New York, NY, USA, July 2021. Association for Computing Machinery. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462618.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3150–3158. PMLR, July 2018.

Deborah Lucas. Measuring the Cost of Bailouts. *Annual Review of Financial Economics*, 11(1): 85–108, 2019.

Christoph Maas. Transportation in graphs and the admittance spectrum. *Discrete Applied Mathematics*, 16(1):31–49, 1987. ISSN 0166218X.

Yanbing Mao, Sadegh Bolouki, and Emrah Akyol. Spread of information with confirmation bias in cyber-social networks. *IEEE Transactions on Network Science and Engineering*, 7(2):688–700, 2018.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the 2020 International Conference on Machine Learning*, 2020.

E. Maskin. A theorem on utilitarianism. *The Review of Economic Studies*, 45(1):93–96, February 1978. ISSN 0034-6527, 1467-937X. doi: 10.2307/2297086.

Gianni Dal Maso. *An Introduction to* Γ*-convergence*. Springer Science & Business Media, December 2012. ISBN 978-1-4612-0327-8.

Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 31(5):1480–1505, 2017.

Antonis Matakos, Sijing Tu, and Aristides Gionis. Tell me something my friends do not know: Diversity maximization in social networks. *Knowledge and Information Systems*, 62(9):3697–3726, 2020.

Robert C Merton. Lifetime Portfolio Selection Under Uncertainty: The Continuous-time Case. *The Review of Economics and Statistics*, pages 247–257, 1969.

Robert C Merton. Optimum Consumption and Portfolio Rules in a Continuous-time Model. *Journal of Economic Theory*, 3:373–413, 1971.

John Stuart Mill. Utilitarianism. In *Seven Masterpieces of Philosophy*. Routledge, 2008. ISBN 978-1-315-50881-8.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. Minimizing Polarization and Disagreement in Social Networks. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, pages 369–378, 2018.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the 2018 ACM Conference on Fairness, Accountability, and Transparency*, 2018.

Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003.

Alejandro Noriega-Campero, Bernardo Garcia-Bulle, Luis Fernando Cantu, Michiel A. Bakker, Luis Tejerina, and Alex Pentland. Algorithmic targeting of social policies: Fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 241–251, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3375784.

Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, September 2017. ISBN 978-0-553-41883-5.

Henri Pagès and Dylan Possamaï. A Mathematical Treatment of Bank Monitoring Incentives. *Finance and Stochastics*, 1(18):39–73, 2014.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax demographic group fairness in federated learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 142–159, June 2022. doi: 10.1145/3531146.3533081.

Christos Papadimitriou. Algorithms, Games, and the Internet. In *Proceedings of the Thirty-third Annual ACM Symposium on Theory of Computing*, pages 749–753, 2001.

ProPublica. COMPAS Recidivism Risk Score Data and Analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

Jiahu Qin, Qichao Ma, Yang Shi, and Long Wang. Recent advances in consensus of multi-agent systems: A brief survey. *IEEE Transactions on Industrial Electronics*, 64(6):4972–4983, 2016.

Miklos Z Racz and Daniel E Rigobon. Towards consensus: Reducing polarization by perturbing social networks. *IEEE Transactions on Network Science and Engineering, forthcoming*, 2023.

Inzamam Rahaman and Patrick Hosein. A model for optimizing article recommendation for reducing polarization. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 107–110, 2021.

Ashesh Rambachan, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. An economic perspective on algorithmic fairness. *AEA Papers and Proceedings*, 110:91–95, May 2020. ISSN 2574-0768, 2574-0776. doi: 10.1257/pandp.20201036.

Ashesh Rambachan, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig. An economic approach to regulating algorithms, 2021.

John Rawls. A Theory of Justice. In *Ethics: Contemporary Readings*. Routledge, 2003. ISBN 978-0-203-49566-7.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Daniel E Rigobon. From utilitarian to rawlsian designs for algorithmic fairness. *arXiv preprint arXiv:2302.03567*, 2023.

Daniel E Rigobon and Ronnie Sircar. Formation of optimal interbank lending networks under liquidity shocks. *arXiv preprint arXiv:2211.12404*, 2022.

Kevin W. S. Roberts. Interpersonal comparability and social choice theory. *The Review of Economic Studies*, 47(2):421, January 1980. ISSN 00346527. doi: 10.2307/2297002.

Jean-Charles Rochet and Jean Tirole. Interbank Lending and Systemic Risk. *Journal of Money, Credit and Banking*, 28(4):733–762, 1996.

L. Christopher G. Rogers. *Optimal Investment*. Springer, 2013.

Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, 3d ed edition, 1976. ISBN 978-0-07-054235-8.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, April 2020.

Fernando P Santos, Yphtach Lelkes, and Simon A Levin. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50), 2021.

Amartya Sen. Welfare inequalities and rawlsian axiomatics. *Theory and Decision*, 7, 1976.

Amartya Sen. Equality of what?, 1979.

Amartya Sen. Commodities and Capabilities. *OUP Catalogue*, 1999.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Huijuan Wang and Piet Van Mieghem. Algebraic connectivity optimization via link addition. In *3d International ICST Conference on Bio-Inspired Models of Network, Information, and Computing Systems*, 2010.

Takamitsu Watanabe and Naoki Masuda. Enhancing the spectral gap of networks by node removal. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 82(4):1–7, 2010.

Zhenhuan Yang, Yan Lok Ko, Kush R. Varshney, and Yiming Ying. Minimax AUC fairness: Efficient algorithm with provable convergence, November 2022.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2): 2473–2480, 2009.

Adam Zawadowski. Entangled Financial Systems. *The Review of Financial Studies*, 26(5):1291–1323, 2013.

Liwang Zhu, Qi Bao, and Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. *Advances in Neural Information Processing Systems*, 34, 2021.