Towards Trustworthy and Data-Driven Social Interventions

by

Aida Rahmattalabi

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(Computer Science)

May 2022

Dedication

To my beloved Amin, my dear parents, and brother

Acknowledgements

I would like to express my deepest gratitude to my advisors, Dr. Phebe Vayanos and Dr. Milind Tambe for their constant support and guidance through the course of my PhD. Thank you so much Phebe for taking me under your guidance. Your investment in my professional and personal development has been truly remarkable and I can not thank you enough for it. Milind, I am very grateful for having you as my mentor. You inspired me to always reach for the best and persevere until I achieve it. Thank you for trusting in me and helping me find my path.

This thesis is the result of close collaboration with other amazing faculty members, researchers and experts inside and outside of USC. I would like to specially thank Dr. Eric Rice for his unwavering support and guidance throughout these years. Eric, I am constantly amazed by the depth of your knowledge and dedication to community-driven research. It has been an absolute pleasure to work with you and learn from you. I would also like to thank my other thesis committee members: Dr. Cyrus Shahabi and Dr. Bistra Dilkina. Thank you so much for your invaluable feedback and mentorship over the years. I also thank Dr. Barman-Adhikari, Dr. Ryan Brown, Dr. Anthony Fulginiti, Dr. David Gray Grant, Daniel Ho (J.D., Ph.D.), Maxwell Izenberg, Dr. Shahin Jabbari, Dr. Ece Kamar, Dr. Hima Lakkaraju and Alice Xiang (J.D.) whom I had the opportunity to work with.

In the last six years, I have had the chance to meet and work with so many great people at USC CAIS. Specially, I would like to thank Sina Aghaei, Elizabeth Bondi, Sarah Cooney, Aaron Ferber, Ben Ford, Shahrzad Gholami, Tye Hines, Qing Jin, Caroline Johnston, Nathan Justin, Debarun Kar, Jackson Killian, Laksh Matai, Aditya Mate, Han Ching Ou, Caleb Robinson,

Aaron Schlenker, Bill Tang, Omkar Thakoor, Kai Wang, Bryan Wilder, Hailey Winetrobe, Lily Xu, Amulya Yadav, Yingxiao Ye and Han Yu. It was absolutely amazing to get to know each and every one of you. I have also been blessed with many great friends who have created a home for me and been there by my side when I most needed it. In particular, I thank Niloufar Zarei who is more than just a friend for me. I thank Haleh Akrami, Arash Fayazi, Zohre Azizi for all the fun days and nights together. I also thank Arman Massahi for being my sport inspirations (and teaching me how to play tennis properly!), Nripsuta Saxena for being such a fun spirit and Ehsan EmamJomeZade, Nami Mogharabin, Shiva Navabi for all the thought-provoking (and fun) conversations. I would also like to specially thank my dear Amin for his love, support and contagious positive attitude. Most importantly, I would like to thank my parents and brother for their encouragement and support throughout my life. I am who I am today because of you.

Funding

My PhD work was supported in part by Smart and Connected Communities program of the National Science Foundation under NSF award No. 1831770 and the US Army Research Office under grant number W911NF1710445.

Table of Contents

List Of Tables

List Of Figures

Abstract

This thesis examines social interventions conducted to address societal challenges such as homelessness, substance abuse or suicide. In most of these applications, it is challenging to purposefully collect data. Hence, we need to rely on social (e.g., social network data) or observational data (e.g., administrative data) to guide our decisions. Problematically, these datasets are prone to different statistical or societal biases. When optimized and evaluated on these data, ostensibly impartial algorithms may result in disparate impacts across different groups. In addition, these domains are plagued by limited resources and/or limited data which create a computational challenge with respect to improving the delivery of these interventions. In this thesis, I investigate the interplay of fairness and these computational challenges which I present in two parts. In the first part, I introduce the problem of fairness in social network-based interventions where I propose to use social network data to enhance interventions that rely on individual's social connectedness such as HIV/suicide prevention or community preparedness against natural disasters. I demonstrate how biases in the social network can manifest as disparate outcomes across groups and describe my approach to mitigate such unfairness. In the second part, I focus on fairness challenges when data is observational. Motivated by the homelessness crisis in the U.S., I study the problem of learning fair resource allocation policies using observational data where I develop a methodology that handles selection bias in the data. I conclude with a critique on the fairness metrics proposed in the literature, both causal and observational (statistical), and I present a novel causal view that addresses the shortcomings of existing approaches. In particular, my findings shed new light on well-known impossibility results from the fair machine learning literature.

Introduction

For long, societies around the globe have struggled with complex societal problems in the areas of social justice and welfare, education or health that disproportionately impact the most vulnerable. Over the years, researchers, practitioners and policymakers have examined a variety of interventions to address these social problems which I refer to as "social interventions." Fueled by recent algorithmic advances, there has been an increasing interest in developing evidence-based, AI-augmented social interventions that have greater reach and impact and are tailored to the needs of affected communities.

In particular, this thesis investigates three social problems that are critical in the current state of our society and aims to develop trustworthy and data-driven algorithmic solutions to address them. First is suicide prevention. Suicide is a critical public health problem in the United States specially among the youth population such as college students, where suicide takes more than a 1000 lives each year [9]. In this regard, the present thesis studies how we can leverage individuals' social support to mitigate the risk of suicidal ideation and death. Another application of this thesis is for landslide risk management. In particular, Sitka, Alaska experiences frequent landslide incidents which cause significant damage and disruption to the lives of those affected by it. Effective risk management depends heavily on timely and reliable access to risk information [144]. In this thesis, I study how we can use social influence to create resilient and informed communities that can protect themselves against landslide. Finally, this thesis investigates solutions to mitigate homelessness. Cities with high homeless population often suffer from shortage

of resources to address this problem. For instance, in Los Angeles County, there are over 63,000 homeless individuals and far fewer housing units to accommodate them. Furthermore, there is a significant disparity in the rate of homelessness across different racial groups, hitting those from minority groups the hardest [78]. To address these problems, this thesis explores equitable and data-driven policies to help match individuals with suitable resources in order to guarantee a high chance of safe and stable exit from homelessness. The problems studied in this thesis are identified through close collaborations with social scientists at RAND Corporation, a nonprofit global policy think tank, and social work academics who specialize in social network science and community-based research.

Figure 1: Project collaborators at RAND Corporation gather comments at the Sitka Sound Science Center in Sitka, Alaska on our research on landslide preparedness.

Despite the wealth of knowledge in fields such as public health, public policy or social work on the underlying social phenomena, transferring that knowledge to develop practical computational models of interventions is non-trivial. Further, these interventions often give rise to highly intractable models which are difficult to optimize. In addition, designing algorithmic solutions in such complex real-world settings are faced with several unique challenges. Below, I highlight some of the challenges that are central to this thesis.

Fairness: Social problems do not affect all groups equally. For instance, most minority groups in the United States experience homelessness at higher rates than Whites, and therefore make up a disproportionate share of the homeless population. African Americans make up 13% of the general population, but more than 40% of the homeless population. Further, in Los Angeles County, recent studies have found racial inequities in outcomes for Black residents of homeless services —particularly residents of Permanent Supportive Housing, a long-term housing intervention —where Black

residents are 39% more likely to return to homelessness than White residents [127]. Similar disparities exist in other areas such as risk of suicide, where studies have identified evidence of widening gaps in rate of suicide across sex, sexual orientation, race/ethnicity, age, and socioeconomic status subgroups among college students [91, 118]. As algorithms enter such socially-sensitive domains, it is critical that they take the welfare of every group and individual into consideration and strive towards equitable outcomes for all.

Data Bias: Data is central to modern decision-making. However, in these settings controlled experiments are typically costly. As a result, most of available data comes from passive observations which are prone to different forms of bias. For example, recent studies have identified structural racism as one of the main factors for the high rate of homelessness among Black people [12]. Such societal biases will inevitably creep into the data that is used to inform the interventions which can be problematic as they may result in algorithms that discriminate against certain individuals or groups, entrenching existing inequalities. There are also naturally occurring biases. For example, it has been shown that individuals have a tendency to associate and bond with similar others, a phenomenon known as *homophily* [124]. While these natural biases are not intrinsically objectionable, care must be taken when using this type of data for various decision-making tasks as it may lead to undesirable disparities in the outcomes of different individuals or groups.

Resource Limitation: Designing social interventions typically involves the allocation of scarce resources, e.g., limited housing units or social worker hours. In Los Angeles County, there are over 63,000 homeless persons and only 21,000 housing units, most of which are temporary housing assistance. In such settings, providers often need to make complex decisions under great uncertainty with small margins of error. This may lead to fraught decisions that either under- or over-serve specific groups. Resource limitation further compounds fairness challenges as it raises the question of who should or should not receive these resources.

Data Scarcity: Real-world settings are permeated by different forms of uncertainty, e.g., unknown availability of intervention participants, that may negatively affect the outcome of these interventions. In practice, there may not be enough data to inform those uncertainties.

This thesis is concerned with tackling the above challenges where a special emphasis is placed on the issue of fairness and its interplay with resource and data limitation. Specifically, the overarching question that the present thesis aims to address is

How can we develop fair, efficient and data-driven algorithms to enhance social interventions?

I investigate this question within the context of the aforementioned social problems, namely suicide prevention, landslide risk management and mitigating homelessness, where I propose computational models of popular interventions as well as equitable and efficient algorithmic solutions. It is noteworthy that the proposed intervention models are not restricted to the above applications and can be generalized to address other problems that share the same underlying characteristics.

Overview of Contributions

This thesis is divided in two parts. The first part introduces the computational problem of fairness in social network-based interventions, i.e., interventions that rely on social support and individuals' connectedness to succeed, such as suicide prevention and community resilience for landslide risk management. This thesis draws on material published in [76, 97, 148, 149, 150, 151, 152, 153].

Chapter 1 focuses on suicide prevention. Gatekeeper training is one of the widely used suicide prevention interventions which involves teaching individuals to recognize and support those in crisis. A successful intervention seeks to achieve a good coverage of the individuals in a social network (e.g., student population). Targeted enlistment of individuals helps achieve more desirable coverage than baseline strategies [77]. However, the performance is significantly affected by the uncertainty in the availability and performance of the training candidates. In collaboration with the schools of social work at the University of Denver and the University of Southern California,

this work proposes a novel intervention model to select a limited number of individuals, with uncertain performance, to identify warning signs of suicide among their peers in a social network. Using social network of youth experiencing homelessness, this work demonstrates how purely coverage-centric algorithms, such as those introduced in [42, 113, 176], may result in discriminatory coverage across different social groups. Devising efficient solutions that perform well across groups, even under worst-case uncertain scenarios, also poses a highly intractable problem. Chapter 1 addresses this problem by providing a novel formulation of the problem as a robust graph covering problem with group fairness constraints. The solution approach is in the form of a tractable approximation applicable to real-world instances. In addition, this work provides a theoretical analysis of price of group fairness (PoF), with and without uncertainty. Specifically, it shows that uncertainty can lead to greater PoF compared to the deterministic case which highlights the trade-off between fairness and robustness. Empirically, the proposed method yields competitive node coverage while significantly improving group fairness over the state-of-the-art methods.

Chapter 2 investigates interventions for community preparedness against natural hazards such as landslide risk. In collaboration with scientists at RAND Corporation and Sitka Sound Science Center, we identified a major challenge associated with landslide risk management to be timely and reliable access to risk information. Community-based interventions can improve risk communication and access to information, particularly in rural and remote contexts. These interventions often seek to engage and educate a limited set of individuals who can act as community-leaders to spread information to others. Algorithmic influence maximization can aid with the choice of "peer leaders" or "influencers" in such interventions. Existing techniques for fair influence maximization require committing to a single fairness measure or are imposed as strict constraints leading to undesirable properties such as wastage of resources [171, 175]. Chapter 2 revisits the problem of fairness in influence maximization from a welfare optimization perspective. It provides a principled characterization of the properties that a fair influence maximization algorithm should satisfy. As a result, it proposes a framework that aggregates the cardinal utilities derived by each

community using isoelastic social welfare functions. Under this framework, the trade-off between fairness and efficiency can be controlled by a single inequality aversion design parameter which is crucial specially when these solutions are deployed at scale. In addition, the proposed framework encompasses as special cases leximin and proportional fairness. It is further shown that the resulting optimization problem is monotone and submodular and can be solved efficiently with optimality guarantees. Extensive experiments on synthetic and real world datasets including a case study on landslide risk management demonstrate the efficacy of the proposed framework.

The second part of this thesis focuses on challenges that arise when data is observational. In this setting, a decision-maker has to rely on passive data observations prone to selection bias. In particular, selection bias occurs when the assignment of individuals in different groups are not completely at random. For instance, individuals who have been exposed to a certain treatment may be systematically different from those who have been assigned to a control group. Similarly, individuals' sensitive attributes may be correlated with other risk factors important for decision-making. One can view this as a selection bias, as individuals in different sensitive groups have different underlying risk distributions. Selection bias poses unique challenges for designing data-driven interventions as well as unfairness evaluation which I explore in Chapters 3 and 4.

Chapter 3 focuses on the problem of mitigating homelessness. Homeless services authorities commonly consider housing as a key solution to homelessness [139]. Despite different government funding programs and services, the number of homeless individuals in the U.S. surpasses the available resources which necessitates strategic allocations to maximize the intervention's effectiveness. A natural, or rather complex, objective for housing allocation is to optimize the expected number of people exiting homelessness from different social groups (e.g., racial groups). However, the treatment effects of different interventions are unknown and heterogeneous. In other words, the likelihood of a successful outcome depends on the joint characteristics of the resource and individual which is unknown to the decision-maker and should be estimated from data. Historical data, on the other hand, suffers from selection bias which poses a challenge for evaluating and

optimizing policies that perform well across different protected groups. In addressing this problem, this work proposes a computational model to match heterogeneous individuals and resources that arrive stochastically over time. Each individual, upon arrival, is assigned to a queue where they wait to be matched to a resource. The resources are assigned in a first come first served (FCFS) fashion according to an eligibility structure that encodes the resource types that serve each queue. This work provides a methodology based on techniques in modern causal inference to construct the individual queues as well as learn the matching outcomes and provide a mixed-integer optimization (MIO) formulation to optimize the eligibility structure. The MIO problem maximizes policy outcome subject to wait time and fairness constraints. It is very flexible, allowing for additional linear domain constraints. Empirical results using data from the U.S. Homeless Management Information System (HMIS) results in wait times as low as an FCFS policy while improving the rate of exit from homelessness for underserved or vulnerable groups (7% higher for the Black individuals and 15% higher for those below 17 years old).

Finally, Chapter 4 studies unfairness evaluation and mitigation in more generic decision-making applications. In recent years, there has been increasing interest in causal reasoning for designing fair decision-making systems due to its compatibility with legal frameworks, interpretability for human stakeholders, and robustness to spurious correlations inherent in observational data, among other factors. The recent attention to causal fairness, however, has been accompanied with great skepticism due to practical and epistemological challenges with applying current causal fairness approaches in the literature. Motivated by the long-standing empirical work on causality in econometrics, social sciences, and biomedical sciences, this work lays out the conditions for appropriate application of causal fairness under the "potential outcomes framework." Specifically, it highlight key aspects of causal inference that are often ignored in the causal fairness literature, namely the importance of specifying the nature and timing of interventions on social categories such as race or gender. Precisely, instead of postulating an intervention on immutable attributes, this work proposes a shift in focus to their perceptions and discuss the

implications for fairness evaluation. Such conceptualization of the intervention is key in evaluating the validity of causal assumptions and conducting sound causal analyses including avoiding post-treatment bias (a form of bias due to variables that have materialized after one's sensitive attribute is observed). Sound application of causal fairness can further address the limitations of existing fairness metrics, including those that depend upon statistical correlations. Specifically, I introduce causal variants of common statistical notions of fairness, and make a novel observation that under the causal framework there is no fundamental disagreement between different notions of fairness. Finally, extensive experiments demonstrate the effectiveness of the proposed approach for evaluating and mitigating unfairness, specially when post-treatment variables are present.

Related Work

Interest in fairness properties of algorithms can be broadly categorized into two themes: fairness in *prediction* and fairness in *decision*. In recent years, there has been an explosion of research focusing on fairness in machine learning (ML). These works aim to ensure that predictions made by ML algorithms are equitable. To this end, different notions of fairness are defined based on one or more sensitive attributes such as age, race or gender [85, 110, 191]. Despite the variety of individual and group fairness definitions, there is still a lack of expressiveness [126]. Most of these definitions focus solely on the inputs and outputs of the algorithm without taking into account the complexities of the downstream task such as constrained allocation or heterogeneity in utility of different individuals or groups [71]. A few exceptions exist in which the authors study a welfare-based prediction model with fairness considerations [54, 86, 92]. It is worth noting that there is a line of work on budgeted ML which considers resource limitations such as computational cost, time or information input [7, 55]. However, these applications do not directly relate to our settings which require resource constraints on the model prediction.

Research on fairness in decision-making and resource allocation, on the other hand, has a long history. Different disciplines, from operations research, computer science, mathematics to

mechanism design and welfare economics, have studied the fair allocation allocation under different assumptions. In this regard, a typical setting concerns a scenario where a central decision-maker must make an allocation of goods to a number of distinct entities (e.g., individuals) in a fair manner. A line of work studies the fair allocation problem among individuals [16, 20, 41], or groups of agents [15, 66, 116, 166] by defining various fairness criteria. The literature tends to focus on several primary notions of fairness: proportional division [174] (every agent receives at least $1/n$ of her perceived value of resources); equitability [73] (every agent equally values their allocations); envy-freeness [180] (every agent values their allocation at least as much as another's) and maximin fairness [156] (the value received by the worse-off agent is maximized).

While these notions capture fairness of allocations in many real-world applications, there are several barriers to their adoption in practice. First, the common assumption in these works is that utilities are given which overlooks the fact that in practice utilities are unknown and predicted utilities, trained on past behavior, are subject to bias. In addition, they assume that individuals' utilities are independent of one another, i.e., changing an individual's utility will not affect other individuals as long as their share of resources is fixed. Moreover, real-world decisions are subject to different forms of uncertainty. Works that study fairness under uncertainty (e.g., unknown demand) [19, 61, 64, 130] often assume full distributional information about the uncertain parameters. In some social settings, however, distributional information is not available and there may be little data to inform our decisions. Finally, fairness/efficiency trade-offs is another crucial consideration that arises in a variety of applications including organ allocation [172] or disaster response [148]. Prior work is often limited to point-solutions, with little quantitative understanding about the trade-off between efficiency and fairness, which impedes the applicability of these solutions. In spite of recent efforts to discover, evaluate and mitigate algorithmic bias and unfairness, data-driven allocation problems continue to form an important topic for fairness considerations, especially as automated systems enter a wide range of application domains far beyond the original computational settings of the problem. As highlighted above, there are

many unresolved challenges that arise when we consider developing these solutions for real-world settings. This thesis focuses on three social domains. However, it is noteworthy that the fairness challenges and the proposed solutions are not restricted to the above applications and can be generalized to other domains that share the same underlying characteristics.

Part I

Fairness in Social Network-Based Interventions

Chapter 1

Robust and Fair Graph Covering

1.1 Introduction

We consider the problem of selecting a subset of nodes (which we refer to as 'monitors') in a graph that can 'cover' their adjacent nodes. We are mainly motivated by settings where monitors are subject to failure and we seek to maximize worst-case node coverage. We refer to this problem as the *robust graph covering*. This problem finds applications in several critical real-world domains, especially in the context of optimizing social interventions on vulnerable populations. Consider for example the problem of designing *Gatekeeper training interventions for suicide prevention*, wherein a small number of individuals can be trained to identify warning signs of suicide among their peers [96]. A similar problem arises in the context of *disaster risk management in remote communities* wherein a moderate number of individuals are recruited in advance and trained to watch out for others in case of natural hazards (e.g., in the event of a landslide [155]). Previous research has shown that social intervention programs of this sort hold great promise [96, 155]. Unfortunately, in these real-world domains, intervention agencies often have very limited resources, e.g., moderate number of social workers to conduct the intervention, small amount of funding to cover the cost of training. This makes it essential to target the right set of monitors to cover a maximum number of nodes in the network. Further, in these interventions, the performance and availability of individuals (monitors) is *unknown* and *unpredictable*. At the same time, robustness is desired to guarantee high coverage even in worst-case settings to make the approach suitable for deployment in the open world.

Robust graph covering problems similar to the one we consider here have been studied in the literature, see e.g., [42, 176]. Yet, a major consideration distinguishes our problem from previous work: namely, the need for fairness. Indeed, when deploying interventions in the open world (especially in sensitive domains impacting life and death like the ones that motivate this work), care must be taken to ensure that algorithms do not discriminate among people with respect to protected characteristics such as race, ethnicity, disability, etc. In other words, we need to ensure

Network Name	Network Size	Worst-case coverage of individuals by racial group (%)				
		White	Black	Hispanic	Mixed	Other
SPY1	95	70	**36**	–	86	94
SPY2	117	78	–	**42**	76	67
SPY3	118	88	–	**33**	95	69
MFP1	165	96	77	69	73	**28**
MFP2	182	**44**	85	70	77	72

Table 1.1: Racial discrimination in node coverage resulting from applying the algorithm in [176] on real-world social networks from two homeless drop-in centers in Los Angeles, CA [17], when 1/3 of nodes (individuals) can be selected as monitors, out of which at most 10% will fail. The numbers correspond to the worst-case percentage of covered nodes across all monitor availability scenarios.

that independently of their group, individuals have a high chance of being covered, a notion we refer to as *group fairness*.

To motivate our approach, consider deploying in the open world a state-of-the art algorithm for robust graph covering (which does not incorporate fairness considerations). Specifically, we apply the solutions provided by the algorithm from [176] on five real-world social networks. The results are summarized in Table 1.1 where, for each network, we report its size and the worst-case coverage by racial group. In all instances, there is significant disparity in coverage across racial groups. As an example, in network SPY1 36% of Black individuals are covered in the worst-case compared to 70% (resp. 86%) of White (resp. Mixed race) individuals. Thus, when maximizing coverage without fairness, (near-)optimal interventions end up mirroring any differences in degree of connectedness of different groups. In particular, well-connected groups at the center of the network are more likely to be covered (protected). Motivated by the desire to support those that are the less well off, we employ ideas from *maximin fairness* to improve coverage of those groups that are least likely to be protected.

We investigate the *robust graph covering problem with fairness constraints*. Formally, given a social network, where each node belongs to a group, we consider the problem of selecting a subset of I nodes (monitors), when at most J of them may fail. When a node is chosen as a monitor and does not fail, all of its neighbors are said to be 'covered' and we use the term 'coverage' to refer

to the total number of covered nodes. Our objective is to maximize worst-case coverage when any J nodes may fail, while ensuring fairness in coverage across groups. We adopt maximin fairness from the Rawlsian theory of justice [156] as our fairness criterion: we aim to maximize the utility of the groups that are worse-off. To the best of our knowledge, ours is the first work enforcing fairness constraints in the context of graph covering subject to node failure.

We make the following contributions: *(i)* We achieve maximin group fairness by incorporating constraints inside a robust optimization model, wherein we require that at least a fraction W of each group is covered, in the worst-case; *(ii)* We propose a novel two-stage robust optimization formulation of the problem for which near-optimal conservative approximations can be obtained as a moderately-sized mixed-integer linear program (MILP). By leveraging the decomposable structure of the resulting MILP, we propose a Benders' decomposition algorithm augmented with symmetry breaking to solve practical problem sizes; *(iii)* We present the first study of price of group fairness (PoF), i.e., the loss in coverage due to fairness constraints in the graph covering problem subject to node failure. We provide upper bounds on the PoF for Stochastic Block Model networks, a widely studied model of networks with community structure; *(iv)* Finally, we demonstrate the effectiveness of our approach on several real-world social networks of homeless youth. Our method yields competitive node coverage while significantly improving group fairness relative to state-of-the-art methods.

1.2 Related Work

This work relates to three streams of literature which we review.

Algorithmic Fairness. With increase in deployments of AI, OR, and ML algorithms for decision and policy-making in the open world has come increased interest in algorithmic fairness. A large portion of this literature is focused on resource allocation systems, see e.g., [33, 111, 192]. Group fairness in particular has been studied in the context of resource allocation problems [53, 165, 173].

A nascent stream of work proposes to impose fairness by means of constraints in an optimization problem, an approach we also follow. This is for example proposed in [4], and in [24, 64], and in [5] for machine learning, resource allocation, and matching problems, respectively. Several authors have studied the price of fairness. In [33], the authors provide bounds for maximin fair optimization problems. Their approach is restricted to convex and compact utility sets. In [21], the authors study price of fairness for indivisible goods with additive utility functions. In our graph covering problem, this property does not hold. Several authors have investigated notions of fairness under uncertainty, see e.g, [18, 72, 130, 192]. These papers all assume full distributional information about the uncertain parameters and cannot be employed in our setting where limited data is available about node availability. Motivated by data scarcity, we take a robust optimization approach to model uncertainty which does not require distributional information. This problem is highly intractable due to the combinatorial nature of both the decision and uncertainty spaces. When fair solutions are hard to compute, "approximately fair" solutions have been considered [111]. In our work, we adopt an approximation scheme. As such, our approach falls under the "approximately fair" category. Recently, several authors have emphasized the importance of fairness when conducting interventions in socially sensitive settings, see e.g., [13, 114, 175]. Our work most closely relates to [175], wherein the authors propose an algorithmic framework for fair influence maximization. We note that, in their work, nodes are not subject to failure and therefore their approach does not apply in our context.

Submodular Optimization. One can view the group-fair maximum coverage problem as a multi-objective optimization problem, with the coverage of each community being a separate objective. In the deterministic case, this problem reduces to the multi-objective submodular optimization problem [48], as coverage has the submodularity (diminishing returns) property. In addition, moderately sized problems of this kind can be solved optimally using integer programming technology. However, when considering uncertainty in node performance/availability, the objective function loses the submodularity property while exact techniques fail to scale to even moderate

problem sizes. Thus, existing (exact or approximate) approaches do not apply. Our work more closely relates to the robust submodular optimization literature. In [42, 142], the authors study the problem of choosing a set of up to I items, out of which J fail (which encompasses as a special case the robust graph covering problem *without* fairness constraints). They propose a greedy algorithm with a constant (0.387) approximation factor, valid for $J = o(\sqrt{I})$, and $J = o(I)$, respectively. Finally, in [176], the authors propose another greedy algorithm with a general bound based on the curvature of the submodular function. These heuristics, although computationally efficient, are coverage-centered and do not take fairness into consideration. Thus, they may lead to discriminatory outcomes, see Table 1.1.

Robust Optimization. Our solution approach closely relates to robust optimization paradigm which is a computationally attractive framework for obtaining equivalent or conservative approximations based on duality theory, see e.g., [23, 33, 189]. Indeed, we show that the robust graph covering problem can be written as a two-stage robust problem with binary second-stage decisions which is highly intractable in general [35]. One stream of work proposes to restrict the functional form of the recourse decisions to functions of benign complexity [32, 36]. Other works rely on partitioning the uncertainty set into finite sets and applying constant decision rules on each partition [36, 38, 84, 146, 182]. The last stream of work investigates the so-called K-adaptability counterpart [30, 47, 84, 154, 181], in which K candidate policies are chosen in the first stage and the best of these policies is selected *after* the uncertain parameters are revealed. Our work most closely relates to [84, 154]. In [84], the authors show that for bounded polyhedral uncertainty sets, linear two-stage robust optimization problems can be approximately reformulated as MILPs. Paper [154] extends this result to a special case of discrete uncertainty sets. We prove that we can leverage this approximation to reformulate robust graph covering problem with fairness constraints *exactly* for a much larger class of discrete uncertainty sets.

1.3 Fair and Robust Graph Covering Problem

We model a social network as a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, in which $\mathcal{N} := \{1, \ldots, N\}$ is the set of all nodes (individuals) and $\mathcal{E}$ is the set of all edges (social ties). A directed edge from ν to n exists, i.e., $(\nu, n) \in \mathcal{E}$, if node n can be covered by ν. We use $\delta(n) := \{\nu \in \mathcal{N} : (\nu, n) \in \mathcal{E}\}$ to denote the set of neighbors (friends) of n in $\mathcal{G}$, i.e., the set of nodes that can cover node n. Each node $n \in \mathcal{N}$ is characterized by a set of attributes (protected characteristics) such as age, race, gender, etc., for which fair treatment is important. Based on these node characteristics, we partition $\mathcal{N}$ into C disjoint groups $\mathcal{N}_c$, $c \in \mathcal{C} := \{1, \ldots, C\}$, such that $\cup_{c \in \mathcal{C}} \mathcal{N}_c = \mathcal{N}$.

We consider the problem of selecting a set of I nodes from $\mathcal{N}$ to act as 'peer-monitors' for their neighbors, given that the availability of each node is unknown a-priori and at most J nodes may fail (be unavailable). We encode the choice of monitors using a binary vector $\boldsymbol{x}$ of dimension N whose nth element is one iff the nth node is chosen. We require $\boldsymbol{x} \in \mathcal{X} := \{\boldsymbol{x} \in \{0,1\}^N : \mathbf{e}^\top \boldsymbol{x} \leq I\}$, where $\mathbf{e}$ is a vector of all ones of appropriate dimension. Accordingly, we encode the (uncertain) node availability using a binary vector $\boldsymbol{\xi}$ of dimension N whose nth element equals one iff node n does not fail (is available). Given that data available to inform the distribution of $\boldsymbol{\xi}$ is typically scarce, we avoid making distributional assumptions on $\boldsymbol{\xi}$. Instead, we view uncertainty as deterministic and set based, in the spirit of robust optimization [23]. Thus, we assume that $\boldsymbol{\xi}$ can take-on any value from the set Ξ which is often referred to as the *uncertainty set* in robust optimization. The set Ξ may for example conveniently capture failure rate information. Thus, we require $\boldsymbol{\xi} \in \Xi := \{\boldsymbol{\xi} \in \{0,1\}^N : \mathbf{e}^\top(\mathbf{e} - \boldsymbol{\xi}) \leq J\}$. A node n is counted as 'covered' if at least one of its neighbors is a monitor and does not fail (is available). We let $\boldsymbol{y}_n(\boldsymbol{x}, \boldsymbol{\xi})$ denote if n is covered for the monitor choice $\boldsymbol{x}$ and node availability $\boldsymbol{\xi}$.

$$\boldsymbol{y}_n(\boldsymbol{x}, \boldsymbol{\xi}) := \mathbb{I}\left(\textstyle\sum_{\nu \in \delta(n)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu \geq 1\right).$$

The coverage is then expressible as $F_{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{\xi}) := \mathbf{e}^\top \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\xi})$. The *robust covering problem* which aims to maximize the worst-case (minimum) coverage under node failures can be written as

$$\max_{\boldsymbol{x} \in \mathcal{X}} \ \min_{\boldsymbol{\xi} \in \Xi} \ F_{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{\xi}). \tag{$\mathcal{RC}$}$$

Problem ($\mathcal{RC}$) ignores fairness and may result in discriminatory coverage with respect to (protected) node attributes , see Table 1.1. We thus propose to augment the robust covering problem with fairness constraints. Specifically, we propose to achieve max-min fairness by imposing fairness constraints on each group's coverage: we require that at least a fraction W of nodes from each group be covered. In [175], the authors show that by conducting a binary search for the largest W for which fairness constraints are satisfied for all groups, the max-min fairness optimization problem is equivalent to the one with fairness constraints. Thus, we write the *robust covering problem with fairness constraints* as

$$\left\{ \max_{\boldsymbol{x} \in \mathcal{X}} \ \min_{\boldsymbol{\xi} \in \Xi} \ \sum_{c \in \mathcal{C}} F_{\mathcal{G},c}(\boldsymbol{x}, \boldsymbol{\xi}) \ : \ F_{\mathcal{G},c}(\boldsymbol{x}, \boldsymbol{\xi}) \geq W |\mathcal{N}_c| \quad \forall c \in \mathcal{C}, \ \forall \boldsymbol{\xi} \in \Xi \right\}, \tag{$\mathcal{RC}_{\text{fair}}$}$$

where $F_{\mathcal{G},c}(\boldsymbol{x}, \boldsymbol{\xi}) := \sum_{n \in \mathcal{N}_c} \boldsymbol{y}_n(\boldsymbol{x}, \boldsymbol{\xi})$ is the coverage of group $c \in \mathcal{C}$. Note that if $|\mathcal{C}| = 1$, Problem ($\mathcal{RC}_{\text{fair}}$) reduces to Problem ($\mathcal{RC}$), and if $\Xi = \{\mathbf{e}\}$, Problem ($\mathcal{RC}_{\text{fair}}$) reduces to the deterministic covering problem with fairness constraints. We emphasize that our approach can handle fairness with respect to more than one protected attribute by either: *(a)* partitioning the network based on joint values of the protected attributes and imposing a max-min fairness constraint for each group; or *(b)* imposing max-min fairness constraints for each protected attribute separately. Problem ($\mathcal{RC}_{\text{fair}}$) is computationally hard due to the combinatorial nature of both the uncertainty and decision spaces. Lemma 1 characterizes its complexity. Proofs of all results are in the supplementary document.

Lemma 1. *Problem ($\mathcal{RC}_{\text{fair}}$) is $\mathcal{NP}$-hard.*

1.4 Price of Group Fairness

In Section 1.3, we proposed a novel formulation of the robust covering problem incorporating fairness constraints, Problem ($\mathcal{RC}_{\text{fair}}$). Unfortunately, adding fairness constraints to Problem ($\mathcal{RC}$) comes at a price to overall worst-case coverage. In this section, we study this *price of group fairness*.

Definition 1. *Given a graph $\mathcal{G}$, the Price of Group Fairness* $\text{PoF}(\mathcal{G}, I, J)$ *is the ratio of the coverage loss due to fairness constraints to the maximum coverage in the absence of fairness constraints, i.e.,*

$$\text{PoF}(\mathcal{G}, I, J) := 1 - \frac{\text{OPT}^{\text{fair}}(\mathcal{G}, I, J)}{\text{OPT}(\mathcal{G}, I, J)}, \tag{1.1}$$

where $\text{OPT}^{\text{fair}}(\mathcal{G}, I, J)$ *and* $\text{OPT}(\mathcal{G}, I, J)$ *denote the optimal objective values of Problems* ($\mathcal{RC}_{\text{fair}}$) *and* ($\mathcal{RC}$)*, respectively, when I monitors can be chosen and at most J of them may fail.*

In this work, we are motivated by applications related to social networks, where it has been observed that people with similar (protected) characteristics tend to interact more frequently with one another, forming friendship groups (communities). This phenomenon, known as *homophily* [125], has been observed for characteristics such as race, gender, education, etc.[56]. This motivates us to study the PoF in Stochastic Block Model (SBM) networks [70], a widely accepted model for networks with community structure. In SBM networks, nodes are partitioned into C disjoint communities $\mathcal{N}_c$, $c \in \mathcal{C}$. Within each community c, an edge between two nodes is present independently with probability p_c^{in}. Between a pair of communities c and $c' \in \mathcal{C}$, edges exist independently with probability $p_{cc'}^{\text{out}}$ and we typically have $p_c^{\text{in}} > p_{cc'}^{\text{out}}$ to capture homophily. Thus, SBM networks are very adequate models for our purpose. We assume w.l.o.g. that the communities are labeled such that: $|\mathcal{N}_1| \leq \ldots \leq |\mathcal{N}_C|$.

1.4.1 Deterministic Case.

We first study the PoF in the deterministic case for which $J = 0$. Lemma 2 shows that there are worst-case networks for which PoF can be arbitrarily bad.

Lemma 2. *Given $\epsilon > 0$, there exists a budget I and a network $\mathcal{G}$ with $N \geq \frac{4}{\epsilon} + 3$ nodes such that* $\mathrm{PoF}(\mathcal{G}, I, 0) \geq 1 - \epsilon$.

Fortunately, as we will see, this pessimistic result is not representative of the networks that are seen in practice. We thus investigate the loss in expected coverage due to fairness constraints, given by

$$\overline{\mathrm{PoF}}(I, J) := 1 - \frac{\mathbb{E}_{\mathcal{G}\sim\mathrm{SBM}}[\mathrm{OPT}^{\mathrm{fair}}(\mathcal{G}, I, J)]}{\mathbb{E}_{\mathcal{G}\sim\mathrm{SBM}}[\mathrm{OPT}(\mathcal{G}, I, J)]}. \tag{1.2}$$

We emphasize that we investigate the loss in the expected coverage rather than the expected PoF for analytical tractability reasons. We make the following assumptions about SBM network.

Assumption 1. *For all communities $c \in \mathcal{C}$, the probability of an edge between two individuals in the community is inversely proportional to the size of the community, i.e., $p_c^{\mathrm{in}} = \Theta(|\mathcal{N}_c|^{-1})$.*

Assumption 2. *For any two communities $c, c' \in \mathcal{C}$, the probability of an edge between two nodes $n \in \mathcal{N}_c$ and $\nu \in \mathcal{N}_{c'}$ is $p_{cc'}^{\mathrm{out}} = O((|\mathcal{N}_c| \log^2 |\mathcal{N}_c|)^{-1})$.*

Assumption 1 is based on the observation that social networks are usually sparse. This means that most individuals do not form too many links, even if the size of the network is very large. Sparsity is characterized in the literature by the number of edges being proportional to the number of nodes which is the direct result of Assumption 1. Assumption 2 is necessary for meaningful community structure in the network. We now present results for the upper bound on $\overline{\mathrm{PoF}}$ in SBM networks.

Proposition 1. *Consider an SBM network model with parameters p_c^{in} and $p_{cc'}^{\text{out}}$, $c, c' \in \mathcal{C}$, satisfying Assumptions 1 and 2. If $I = O(\log N)$, then*

$$\overline{\text{PoF}}(I, 0) = 1 - \frac{\sum_{c\in\mathcal{C}} |\mathcal{N}_c|}{\sum_{c\in\mathcal{C}} |\mathcal{N}_c| d(C)/d(c)} - o(1), \text{ where } d(c) := \log |\mathcal{N}_c| (\log\log |\mathcal{N}_c|)^{-1}.$$

Proof Sketch. First, we show that under Assumption 1, the coverage within each community is the sum of the degrees of the monitoring nodes. Then, using the assumption on I in the premise of the proposition (which can be interpreted as a "small budget assumption"), we evaluate the maximum coverage within each community. Next, we show that between-community coverage is negligible compared to within-community coverage. Thus, we determine the distribution of the monitors, in the presence and absence of fairness constraints. PoF is computed based on the these two quantities. ■

1.4.2 Uncertain Case.

Here, imposing fairness is more challenging as we do not know a-priori which nodes may fail. Thus, we must ensure that fairness constraints are satisfied under all failure scenarios.

Proposition 2. *Consider an SBM network model with parameters p_c^{in} and $p_{cc'}^{\text{out}}$, $c, c' \in \mathcal{C}$, satisfying Assumptions 1 and 2. If $I = O(\log N)$, then*

$$\overline{\text{PoF}}(I, J) = 1 - \frac{\eta \sum_{c\in\mathcal{C}} |\mathcal{N}_c|}{(I - J) \times d(C)} - \frac{J \sum_{c\in\mathcal{C}\setminus\{C\}} d(c)}{(I - J) \times d(C)} - o(1),$$

where $d(c)$ is as in Proposition 1 and $\eta := (I - CJ) \left(\sum_{c\in\mathcal{C}} |\mathcal{N}_c|/d(c)\right)^{-1}$.

Proof Sketch. The steps of the proof are similar to those in the proof of Proposition 1 with the difference that, under uncertainty, monitors should be distributed such that the fairness constraints are satisfied even after J nodes fail. Thus, we quantify a minimum number of monitors

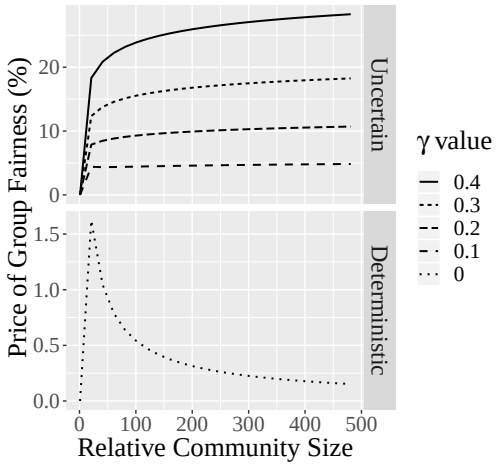


Figure 1.1: PoF in the uncertain (top) and deterministic (bottom) settings for SBM networks consisting of two communities ($\mathcal{C} = \{1, 2\}$) where the size of the first community is fixed at $|\mathcal{N}_1| = 20$ and the size of the other community is increased from $|\mathcal{N}_2| = 20$ to $10,000$. In the uncertain setting, γ denotes the fraction of nodes that fail.

that should be allocated to each community. We then determine the worst-case coverage both in the presence and absence of fairness constraints. PoF is computed based on these two quantities.

■

Propositions 1 and 2 show how PoF changes with the relative sizes of the communities for the deterministic and uncertain cases, respectively. Our analysis shows that without fairness, one should place all the monitors in the biggest community. Under a fair allocation however monitors are more evenly distributed (although larger communities still receive a bigger share). Figure 1.1 illustrates the PoF results in the case of two communities for different failure rates γ ($J = \gamma I$), ignoring the $o(.)$ order terms. We keep the size of the first (smaller) community fixed and vary the size of the larger community. In both cases, if $|\mathcal{N}_1| = |\mathcal{N}_2|$, the PoF is zero since uniform distribution of monitors is optimal. As $|\mathcal{N}_2|$ increases, the PoF increases in both cases. Further increases in $|\mathcal{N}_2|$ result in a decrease in the PoF for the deterministic case: under a fair allocation, the bigger community receives a higher share of monitors which is aligned with the total coverage objective. Under uncertainty however, the PoF is non-decreasing: to guarantee

fairness, additional monitors must be allocated to the smaller groups. This also explains why PoF increases with γ.

1.5 Solution Approach

Given the intractability of Problem ($\mathcal{RC}_{\text{fair}}$), see Lemma 1, we adopt a conservative approximation approach. To this end, we proceed in three steps. First, we note that a difficulty of Problem ($\mathcal{RC}_{\text{fair}}$) is the discontinuity of its objective function. Thus, we show that ($\mathcal{RC}_{\text{fair}}$) can be formulated equivalently as a *two-stage* robust optimization problem by introducing a *fictitious* counting phase *after* $\boldsymbol{\xi}$ is revealed. Second, we propose to approximate this decision made in the counting phase (which decides, for each node, whether it is or not covered). Finally, we demonstrate that the resulting approximate problem can be formulated equivalently as a moderately sized MILP, wherein the trade-off between suboptimality and tractability can be controlled by a single design parameter.

1.5.1 Equivalent Reformulation.

For any given choice of $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \Xi$, the objective $F_{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{\xi})$ can be explicitly expressed as the optimal objective value of a covering problem. As a result, we can express ($\mathcal{RC}_{\text{fair}}$) equivalently as the two-stage *linear* robust problem

$$\max_{\boldsymbol{x}\in\mathcal{X}} \min_{\boldsymbol{\xi}\in\Xi} \max_{\boldsymbol{y}\in\mathcal{Y}} \left\{ \sum_{n\in\mathcal{N}} \boldsymbol{y}_n \ : \boldsymbol{y}_n \leq \sum_{\nu\in\delta(n)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu, \ \forall n \in \mathcal{N} \right\}, \tag{1.3}$$

see Proposition 3 below. The second-stage binary decision variables $\boldsymbol{y} \in \mathcal{Y} := \{\boldsymbol{y} \in \{0,1\}^N : \sum_{n\in\mathcal{N}_c} \boldsymbol{y}_n \geq W|\mathcal{N}_c|, \ \forall c \in \mathcal{C}\}$ admit a very natural interpretation: at an optimal solution, $\boldsymbol{y}_n = 1$ if and only if node n is covered. Henceforth, we refer to $\boldsymbol{y}$ as a *covering scheme*.

Definition 2 (Upward Closed Set)**.** *A set $\mathcal{X}$ given as a subset of the partially ordered set $[0,1]^N$ equipped with the element-wise inequality, is said to be upward closed if for all $\boldsymbol{x} \in \mathcal{X}$ and $\bar{\boldsymbol{x}} \in [0,1]^N$ such that $\bar{\boldsymbol{x}} \geq \boldsymbol{x}$, it holds that $\bar{\boldsymbol{x}} \in \mathcal{X}$.*

Intuitively, sets involving lower bound constraints on the (sums of) parameters satisfy this definition. For example, sets that require a minimum fraction of nodes to be available. We can also consider group-based availability and require a minimum fraction of nodes to be available in every group.

Assumption 3. *We assume that: The set Ξ is defined through $\Xi := \{0,1\}^N \cap \mathcal{T}$ for some upward closed set $\mathcal{T}$ given by $\mathcal{T} := \{\boldsymbol{\xi} \in \mathbb{R}^N : \boldsymbol{A}\boldsymbol{\xi} \geq \boldsymbol{b}\}$, with $\boldsymbol{A} \in \mathbb{R}^{R \times N}$ and $\boldsymbol{b} \in \mathbb{R}^R$.*

Proposition 3. *Problems ($\mathcal{RC}_{\text{fair}}$) and (1.3) are equivalent.*

***K*-adaptability Counterpart.** Problem (1.3) has the advantage of being linear. Yet, its max-min-max structure precludes us from solving it directly. We investigate a conservative approximation to Problem (1.3) referred to as *K-adaptability counterpart*, wherein K *candidate* covering schemes are chosen in the first stage and the best (feasible and most accurate) of those candidates is selected after $\boldsymbol{\xi}$ is revealed. Formally, the K-adaptability counterpart of Problem (1.3) is

$$\underset{\substack{\boldsymbol{x} \in \mathcal{X} \\ \boldsymbol{y}^k \in \mathcal{Y},\, k \in \mathcal{K}}}{\text{maximize}} \quad \min_{\boldsymbol{\xi} \in \Xi} \max_{k \in \mathcal{K}} \left\{ \sum_{n \in \mathcal{N}} \boldsymbol{y}_n^k \;:\; \boldsymbol{y}_n^k \leq \sum_{\nu \in \delta(n)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu \;\; \forall n \in \mathcal{N} \right\}, \tag{1.4}$$

where $\boldsymbol{y}^k$ denotes the kth candidate covering scheme, $k \in \mathcal{K}$. We emphasize that the covering schemes are not inputs but rather *decision variables* of the K-adaptability problem. Only the value K is an input. The optimization problem will identify the best K covering schemes that satisfy all the constraints including fairness constraints. The trade-off between optimality and computational complexity of Problem (1.4) can conveniently be tuned using the single parameter K.

Reformulation as an MILP. We derive an exact reformulation for the K-adaptability counterpart (1.4) of the *robust covering problem* as a moderately sized MILP. Our method extends the results from [154] to significantly more general uncertainty sets that are useful in practice, and to problems involving constraints on the set of covered nodes. Henceforth, we let $\mathcal{L} := \{0, \ldots, N\}^K$, and we define $\mathcal{L}_+ := \{\boldsymbol{\ell} \in \mathcal{L} : \boldsymbol{\ell} > \mathbf{0}\}$ and $\mathcal{L}_0 := \{\boldsymbol{\ell} \in \mathcal{L} : \boldsymbol{\ell} \not> \mathbf{0}\}$. We present a variant of the generic K-adaptability Problem (1.4), where the uncertainty set Ξ is parameterized by vectors $\boldsymbol{\ell} \in \mathcal{L}$. Each $\boldsymbol{\ell}$ is a K-dimensional vector, whose kth component encodes if the kth covering scheme satisfies the constraints of the second stage maximization problem. In this case, $\boldsymbol{\ell}_k = 0$. Else, if the kth covering scheme is infeasible, $\boldsymbol{\ell}_k$ is equal to the index of a constraint that is violated.

Theorem 1. *Under Assumption 3, Problem* (1.4) *is equivalent to the mixed-integer bilinear program*

$$
\begin{aligned}
\max \quad & \tau \\
\text{s.t.} \quad & \tau \in \mathbb{R},\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{y}^k \in \mathcal{Y}\ \ \forall k \in \mathcal{K} \\
& \left.\begin{aligned}
& \boldsymbol{\theta}(\boldsymbol{\ell}),\ \boldsymbol{\beta}^k(\boldsymbol{\ell}) \in \mathbb{R}^N_+,\ \boldsymbol{\alpha}(\boldsymbol{\ell}) \in \mathbb{R}^R_+,\ \boldsymbol{\nu}(\boldsymbol{\ell}) \in \mathbb{R}^K_+,\ \boldsymbol{\lambda}(\boldsymbol{\ell}) \in \Delta_K(\boldsymbol{\ell}) \\
& \tau \ \leq\ -\mathbf{e}^\top \boldsymbol{\theta}(\boldsymbol{\ell}) + \boldsymbol{\alpha}(\boldsymbol{\ell})^\top \boldsymbol{b} - \sum_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k \neq 0}} \left(\boldsymbol{y}^k_{\boldsymbol{\ell}_k} - 1\right) \boldsymbol{\nu}_k(\boldsymbol{\ell}) + \dots \\
& \qquad \dots + \sum_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k = 0}} \sum_{n \in \mathcal{N}} \boldsymbol{y}^k_n \boldsymbol{\beta}^k_n(\boldsymbol{\ell}) + \sum_{k \in \mathcal{K}} \boldsymbol{\lambda}_k(\boldsymbol{\ell}) \sum_{n \in \mathcal{N}} \boldsymbol{y}^k_n \\
& \boldsymbol{\theta}_n(\boldsymbol{\ell}) \ \leq\ \boldsymbol{A}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) + \sum_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k \neq 0}} \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{x}_\nu \boldsymbol{\nu}_k(\boldsymbol{\ell}) - \sum_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k = 0}} \sum_{\nu \in \delta(n)} \boldsymbol{x}_\nu \boldsymbol{\beta}^k_n(\boldsymbol{\ell})\ \ \forall n \in \mathcal{N}
\end{aligned}\right\} \forall \boldsymbol{\ell} \in \mathcal{L}_0 \\
& \left.\begin{aligned}
& \boldsymbol{\theta}(\boldsymbol{\ell}) \in \mathbb{R}^N_+,\ \boldsymbol{\alpha}(\boldsymbol{\ell}) \in \mathbb{R}^R_+,\ \boldsymbol{\nu}(\boldsymbol{\ell}) \in \mathbb{R}^K_+ \\
& 1 \leq -\mathbf{e}^\top \boldsymbol{\theta}(\boldsymbol{\ell}) + \boldsymbol{\alpha}(\boldsymbol{\ell})^\top \boldsymbol{b} - \sum_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k \neq 0}} \left(\boldsymbol{y}^k_{\boldsymbol{\ell}_k} - 1\right) \boldsymbol{\nu}_k(\boldsymbol{\ell}) \\
& \boldsymbol{\theta}_n(\boldsymbol{\ell}) \ \leq\ \boldsymbol{A}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) + \sum_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k \neq 0}} \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{x}_\nu \boldsymbol{\nu}_k(\boldsymbol{\ell})\ \ \forall n \in \mathcal{N}
\end{aligned}\right\} \forall \boldsymbol{\ell} \in \mathcal{L}_+,
\end{aligned}
\tag{1.5}
$$

which can be reformulated equivalently as an MILP using standard "Big-M" techniques since all bilinear terms are products continuous and binary variables. The size of this MILP scales with $|\mathcal{L}| = (N+1)^K$*; it is polynomial in all problem inputs for any fixed* K*.*

Proof Sketch. The reformulation relies on three key steps: First, we partition the uncertainty set by using the parameter $\boldsymbol{\ell}$. Next, we show that by relaxing the integrality constraint on the uncertain parameters $\boldsymbol{\xi}$, the problem remains unchanged. This is the key result that enables us to provide an equivalent formulation for Problem (1.4). Finally, we employ linear programming

duality theory, to reformulate the robust optimization formulation over each subset. As a result, the formulation has two sets of decision variable: *(a)* The decision variables of the original problem; *(b)* Dual variables parameterized by $\boldsymbol{\ell}$ which emerge from the dualization. ■

1.5.2 Bender's Decomposition.

In Problem (1.5), once binary variables $\boldsymbol{x}$ and $\{\boldsymbol{y}^k\}_{k\in\mathcal{K}}$ are fixed, the problem decomposes across $\boldsymbol{\ell}$, i.e., all remaining variables are real valued and can be found by solving a linear program for each $\boldsymbol{\ell}$. Bender's decomposition is an *exact* solution technique that leverages such decomposable structure for more efficient solution [25, 37]. Each iteration of the algorithm starts with the solution of a relaxed master problem, which is fed into the subproblems to identify violated constraints to add to the master problem. The process repeats until no more violated constraints can be identified. The formulations of master and subproblems are provided in Appendix A.

Symmetry Breaking Constraints. Problem (1.5) presents a large amount of symmetry. Indeed, given K candidate covering schemes $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^K$, their indices can be permuted to yield another, distinct, feasible solution with identical cost. The symmetry results in significant slow down of the Brand-and-Bound procedure [39]. Thus, we introduce symmetry breaking constraints in the formulation (1.5) that stipulate the candidate covering schemes be lexicographically decreasing. We refer to [181] for details.

1.6 Results on Social Networks of Homeless Youth

We evaluate our approach on the five social networks from Table 1.1. Details on the data are provided in Section A.1. We investigate the robust graph covering problem with maximin racial fairness constraints. All experiments were ran on a Linux 16GB RAM machine with Gurobi v6.5.0.

First, we compare the performance of our approach against the greedy algorithm of [176] and the degree centrality heuristic (DC). The results are summarized in Figure 1.2 (left). From the

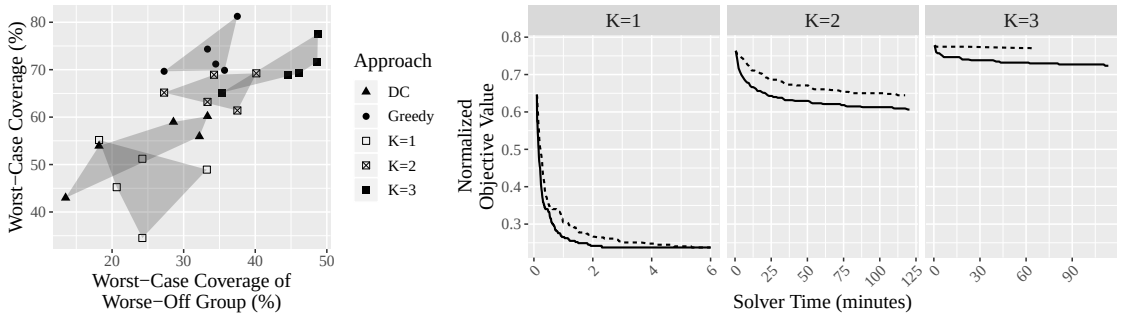


Figure 1.2: Left figure: Solution quality (overall worst-case coverage versus worst-case coverage of the group that is worse-off) for each approach (DC, Greedy, and K-adaptability for $K = 1, 2, 3$); The points represent the results of each approach applied to each of the five real-world social networks from Table 1.1; Each shaded area corresponds to the convex hull of the results associated with each approach; Approaches that are more fair (resp. efficient) are situated in the right- (resp. top-)most part of the graph. Right figure: Average of the ratio of the objective value of the master problem to the network size (across the five instances) in dependence of solver time for the Bender's decomposition approach (dotted line) and the Bender's decomposition approach augmented with symmetry breaking constraints (solid line). For both sets of experiments, the setting was $I = N/3$ and $J = 3$.

figure, we observe that an increase in K results in an increase in performance along both axes, with a significant jump from $K = 1$ to $K = 2, 3$ (recall that K controls complexity/optimality trade-off of our approximation). We note that the gain starts diminishing from $K = 2$ to $K = 3$. Thus, we only run up to $K = 3$. In addition the computational complexity of the problem increases exponentially with K, limiting us to increase K beyond 3 for the considered instances. As demonstrated by our results, $K \sim 3$ was sufficient to considerably improve fairness of the covering at moderate price to efficiency. Compared to the baselines, with $K = 3$, we significantly improve the coverage of the worse-off group over greedy (resp. DC) by 11% (resp. 23%) on average across the five instances.

Second, we investigate the effect of uncertainty on the coverage of the worse-off group and on the PoF, for both the deterministic ($J = 0$) and uncertain ($J > 0$) cases as the number of monitors I is varied in the set $\{N/3, N/5, N/7\}$. These settings are motivated by numbers seen in practice (typically, the number of people that can be invited is 15-20% of network size). Our results are summarized in Table 1.2. Indeed, from the table, we see for example that for $I = N/3$

Name	Size N	Improvement in Min. Percentage Covered (%)						PoF (%)					
		Uncertainty Level J						Uncertainty Level J					
		0	1	2	3	4	5	0	1	2	3	4	5
SPY1	95	15	16	14	10	10	9	1.4	1.0	2.1	1.3	3.3	4.2
SPY2	117	20	14	9	10	8	10	0.0	1.2	3.7	3.3	3.6	3.7
SPY3	118	20	16	16	15	11	10	0.0	3.4	4.8	6.4	3.2	4.0
MFP1	165	17	15	7	11	14	9	0.0	3.1	5.4	2.4	6.3	4.4
MFP2	182	11	12	10	9	12	12	0.0	1.0	1.0	2.2	2.4	3.6
Avg. ($I = N/3$)		16.6	14.6	11.2	11.0	11.0	10.0	0.3	1.9	3.4	3.1	3.8	4.0
Avg. ($I = N/5$)		15.0	13.8	14.0	10.0	9.0	6.7	0.6	2.1	3.2	3.2	3.9	3.8
Avg. ($I = N/7$)		12.2	11.4	11.2	11.4	8.2	6.4	0.1	2.5	3.5	3.2	3.5	4.0

Table 1.2: Improvement on the worst-case coverage of the worse-off group and associated PoF for each of the five real-world social networks from Table 1.1. The first five rows correspond to the setting $I = N/3$. In the interest of space, we only show averages for the settings $I = N/5$ and $I = N/7$. In the deterministic case ($J = 0$), the PoF is measured relative the coverage of the true optimal solution (obtained by solving the integer programming formulation of the graph covering problem). In the uncertain case ($J > 0$), the PoF is measured relative to the coverage of the greedy heuristic of [176].

and $J = 0$ our approach is able to improve the coverage of the worse-off group by 11-20% and for $J > 0$ the improvement in the worse-case coverage of the worse-off group is 7-16%. On the other hand, the PoF is very small: 0.3% on average for the deterministic case and at most 6.4% for the uncertain case. These results are consistent across the range of parameters studied. We note that the PoF numbers also match our analytical results on PoF in that uncertainty generally induces higher PoF.

Third, we perform a head-to-head comparison of our approach for $K = 3$ with the results in Table 1.1. Our findings are summarized in Table A.3 in Section A.1. As an illustration, in SPY3, the worst-case coverage by racial group under our approach is: White 90%, Hispanic 44%, Mixed 85% and Other 87%. These numbers suggest that coverage of Hispanics (the worse-off group) has increased from 33% to 44%, a significant improvement in fairness. To quantify the overall loss due to fairness, we also compute PoF values. The maximum PoF across all instances was at most 4.2%, see Table A.3.

Finally, we investigate the benefits of augmenting our formulation with symmetry breaking constraints. Thus, we solve all five instances of our problem with the Bender's decomposition

approach with and without symmetry breaking constraints. The results are summarized in Figure 1.2 (right). Across our experiments, we set a time limit of 2 hours since little improvement was seen beyond that. In all cases, and in particular for $K = 2$ and 3, symmetry breaking results in significant speed-ups. For $K = 3$ (and contrary to Bender's decomposition augmented with symmetry breaking), Bender's decomposition alone fails to solve the master problem to optimality within the time limit. We would like to remark that employing K-adaptability is necessary: indeed, Problem ($\mathcal{RC}_{\text{fair}}$) would not fit in memory. Similarly, using Bender's decomposition is needed: even for moderate values of K (2 to 3), the K-adaptability MILP (1.5) could not be loaded in memory.

1.7 Conclusion and Broader Impact

We believe that the robust graph covering problem with fairness constraints is worthwhile to investigate. It poses a huge number of challenges and holds great promise in terms of the realm of possible real-world applications with important potential societal benefits, e.g., to prevent suicidal ideation and death and to protect individuals during disasters such as landslides.

Chapter 2

Fair Influence Maximization via Welfare Optimization

2.1 Introduction

The success of many behavioral, social, and public health interventions relies heavily on effectively leveraging social networks [96, 175, 178]. For instance, health interventions such as suicide/HIV prevention [190] and community preparedness against natural disasters involve finding a small set of well-connected individuals who can act as peer-leaders to detect warning signals (suicide prevention) or disseminate relevant information (HIV prevention or landslide risk management). The influence maximization framework has been employed to find such individuals [185]. However, such interventions may lead to discriminatory solutions as individuals from racial minorities or LGBTQ communities may be disproportionately excluded from the benefits of the intervention [151, 175].

Recent work has incorporated fairness directly into influence maximization by proposing various notions of fairness such as maximin fairness [151] and diversity constraints [175]. Maximin fairness aims at improving the minimum amount of influence that any community receives. Inspired by the game theory literature, diversity constraints ensure that each community is at least as well-off had they received their share of resources proportional to their size and allocated them internally. Each of these notions offers a unique perspective on fairness. However, they also come with drawbacks. For example maximin fairness can result in significant degradation in total influence due to its stringent requirement to help the worst-off group as much as possible, where in reality it may be hard to spread the influence to some communities due to their sparse connections. On the other hand, while the diversity constraints aim at taking the community's ability in spreading influence into account, it does not explicitly account for reducing inequality (i.e., does not exhibit *inequality aversion*). Consequently, there is no universal agreement on what fairness means and in fact, it is widely known that fairness is domain dependent [135]. For example, excluding vulnerable communities from suicide prevention might have higher negative consequences compared to interventions promoting a healthier lifestyle.

Building on cardinal social welfare theory from the economics literature and principles of social welfare, we propose a principled characterization of the properties of social influence maximization solutions. In particular, we propose a framework for fair influence maximization based on social welfare theory, wherein the cardinal utilities derived by each community are aggregated using the isoelastic social welfare functions [26]. Isoelastic functions are in the general form of $u^\alpha/\alpha, \alpha < 1, \alpha \neq 0$ and $\log u, \alpha = 0$ where α is a constant and controls the aversion to inequality and u is the utility value. They are used to measure the *goodness* or *desirability* of a utility distribution. However, due to the structural dependencies induced by the underlying social network, i.e., between-community and within-community edges, social welfare principles cannot be directly applied to our problem. Our contributions are as follows:

- We extend the cardinal social welfare principles including the *transfer principle* to the influence maximization framework, which is otherwise not applicable. We also propose a new principle which we call *utility gap reduction*. This principle aims to avoid situations where high aversion to inequality leads to even more utility gap, caused by between-community influence spread.

- We generalize the theory regarding these principles and show that for all problem instances, there does not exist a welfare function that satisfies all principles. Nevertheless, we show that if all communities are disconnected from one another (no between-community edges), isoelastic welfare functions satisfy all principles. This result highlights the importance of network structure, specifically between-community edges.

- Under this framework, the trade-off between fairness and efficiency can be controlled by a single *inequality aversion* parameter α. This allows a decision-maker to effectively trade-off quantities like utility gap and total influence by varying this parameter in the welfare function. We then incorporate these welfare functions as objective into an optimization problem

to rule out undesirable solutions. We show that the resulting optimization problem is monotone and submodular and, hence, can be solved with a greedy algorithm with optimality guarantees; *(iv)* Finally, we carry out detailed experimental analysis on synthetic and real social networks to study the trade-off between total influence spread and utility gap. In particular, we conduct a case study on the social network-based landslide risk management in Sitka, Alaska. We show that by choosing α appropriately we can flexibly control utility gap (4%-26%) and the resulting influence degradation (36% - 5%).

2.2 Related Work

Artificial Intelligence and machine learning algorithms hold great promise in addressing many pressing societal problems. These problems often pose complex ethical and fairness issues which need to be addressed before the algorithms can be deployed in the real world. The nascent field of algorithmic fairness has emerged to address these fairness concerns. To this end, different notions of fairness are defined based on one or more *sensitive attributes* such as age, race or gender. For example, in the classification and regression setting, these notions mainly aim at equalizing a statistical quantity across different communities or populations [85, 191]. While surveying the entirety of this field is out of the score scope (see e.g., [28] for a recent survey), we point out that there is a wide range of fairness notions defined across different settings and it has been shown that the right notion is problem dependent [27, 135] and also different notions of fairness can be incompatible with each other [110]. Thus, care must be taken when we employ these notions of fairness across different applications.

Motivated by the importance of fairness when conducting interventions in social initiatives, fair influence maximization has received a lot of attention recently [6, 72, 151, 175]. These works have incorporated fairness directly into the influence maximization framework by (1) relying on either Rawlsian theory of justice [156, 151], (2) game theoretic principles [175] or (3) equality based

notions [6, 171]. We will discuss the first two approaches in more details in Sections 2.4 and 2.5, as well as in our experimental section. Equality based approaches strive for equal outcomes across different communities. In general, strict equality is hard to achieve and may lead to wastage of resources. This is amplified in influence maximization as different communities have different capacities in being influenced (e.g., marginalized communities are hard to reach). In [72], the authors investigate the notion of information access gap, where they propose maximizing the minimum probability that an individual is being influenced/informed to constrain this gap. As a result they study fairness at an individual level while we study fairness at the group level. Also, their notion of access gap is limited to the gap in a bipartition of the network which is in principle different from utility gap that we study which accommodates arbitrary number of protected groups. Similar to our work, in [6] the authors also study utility gap. They propose an optimization model that directly penalizes utility gap which they solve via a surrogate objective function. Their surrogate functions are in the form of a sum of concave functions of the group utilities which are aggregated with arbitrary weights. Unlike their work, our approach takes an axiomatic approach with strong theoretical justifications and it does not allow for arbitrary concave functions and weights as they violate the welfare principles.

There has also been a long line of work considering fairness in resource allocation problems (see e.g., [33, 111, 43, 41]). More recently, group fairness has been studied in the context of resource allocation problems [53, 64, 24] and specifically in graph covering problems [151]. In resource allocation setting, *maximin fairness* and *proportional fairness* are widely adopted fairness notions. Proportional fairness is a notion introduced for bandwidth allocation [43]. An allocation is proportionally fair if the sum of percentage-wise changes in the utilities of all groups cannot be improved with another allocation. In classical resource allocation problems, each individual or group has a utility function that is independent of the utilities of others individuals or groups. However, this is not the case in influence maximization due to the underlying social network structure i.e., the between-community edges which makes our problem distinct from the classical

resource allocation problems. We note that, while in the bandwidth allocation setting there is also a network structure, the utility of each vertex is still independent of the other vertices and is only a function of the amount of resources that the vertex receives.

Social welfare functions have been used within the economic literature to study trade-offs between equality and efficiency [167] and have been widely adopted in different decision making areas including health [1]. Recently, the authors in [86] proposed to study inequality aversion and welfare through cardinal welfare theory in the context of regression problems. Their main contribution is to use this theory to draw attention to other fairness considerations beyond equality. However, the classical social welfare theory, does not readily extend to our setting due to dependencies induced by the between-community connections. Indeed, extending those principles is a contribution of our work.

2.3 Problem Formulation

We use $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to denote a graph (or network) in which $\mathcal{V}$ is the set of N vertices and $\mathcal{E}$ is the set of edges. In the *influence maximization problem*, a decision-maker chooses a set of at most K vertices to influence (or activate). The selected vertices then spread the influence in rounds according to the *Independent Cascade Model* [105].[1] Under this model, each newly activated vertex spreads the influence to its neighbors independently and with a fixed probability $p \in [0, 1]$. The process continues until no new vertices are influenced. We use $\mathcal{A}$ to denote the initial set of vertices, also referred to as influencer vertices. The goal of the decision-maker is to select a set $\mathcal{A}$ to maximize the expected number of vertices that are influenced at the end of this process. Each vertex of the graph belongs to one of the disjoint communities (empty intersection) $c \in \mathcal{C} := \{1, \ldots, C\}$ such that $\mathcal{V}_1 \cup \cdots \cup \mathcal{V}_C = \mathcal{V}$ where $\mathcal{V}_c$ denotes the set of vertices that belong to community c. This partitioning can be induced by, e.g., the intersection of a set of (protected)

[1] Our framework is also applicable to other forms of diffusion such as Linear Threshold Model [105]

attributes such as race or gender for which fair treatment is important. We use N_c to denote the size of community c, i.e., $N_c = |\mathcal{V}_c|$. Furthermore, communities may be disconnected, in which case $\forall c, c' \in \mathcal{C}$ and $\forall v \in \mathcal{V}_c, v' \in \mathcal{V}_{c'}$, there is no edge between v and v' (i.e., $(v, v') \notin \mathcal{E}$). We define $\mathcal{A}^\star := \{\mathcal{A} \subseteq \mathcal{V} \mid |\mathcal{A}| \leq K\}$ as the set of budget-feasible influencers. Finally, for any choice of influencers $\mathcal{A} \in \mathcal{A}^\star$, we let $u_c(\mathcal{A})$ denote the utility, i.e., the expected fraction of the influenced vertices of community c, where the expectation is taken over randomness in the spread of influence. The standard influence maximization problem solves the optimization problem

$$\underset{\mathcal{A} \in \mathcal{A}^\star}{\text{maximize}} \sum_{c \in \mathcal{C}} N_c u_c(\mathcal{A}). \tag{2.1}$$

When clear from the context, we will drop the dependency of $u_c(\mathcal{A})$ on $\mathcal{A}$ to minimize notational overhead.

2.4 Existing Notions of Fairness

Problem (2.1) solely attempts to maximize the total influence which is also known as the *utilitarian* approach. Existing fair influence maximization problems are variants of Problem (2.1) involving additional constraints. We detail these below.

Maximin Fairness (MMF). Based on the Rawlsian theory [156], MMF [175] aims to maximize the utility of the worst-off community. Precisely, MMF only allows $\mathcal{A} \in \mathcal{A}^\star$ that satisfy the following constraint

$$\min_{c \in \mathcal{C}} \; u_c(\mathcal{A}) \geq \gamma, \quad \text{where } \mathcal{A} \in \mathcal{A}^\star,$$

where the left term is the utility of the worst-off community and γ is the highest value for which the constraint is feasible.

Diversity Constraints (DC). Inspired by the game theoretic notion of core, DC requires that every community obtains a utility higher than when it receives resources proportional to its

size and allocates them internally [175]. This is illustrated by the following constraint where U_c denotes the maximum utility that community c can achieve with a budget equal to $\lfloor KN_c/N \rfloor$.

$$u_c(\mathcal{A}) \geq U_c, \quad \forall c \in \mathcal{C} \text{ where } \mathcal{A} \in \mathcal{A}^\star. \tag{2.2}$$

DC sets utility lower bounds for communities based on their relative sizes and how well they can spread influence internally. As a result, it does not explicitly account for reducing inequalities and may lead to high influence gap. We show this both theoretically and empirically in Sections 2.5.4 and 2.6.

Demographic Parity (DP). Formalizing the legal doctrine of disparate impact [191], DP requires the utility of all communities to be roughly the same. For any $\delta \in [0, 1)$, DP implies the constraints [4, 6, 171]

$$\left|u_c(\mathcal{A}) - u_{c'}(\mathcal{A})\right| \leq \delta, \;\; \forall c, c' \in \mathcal{C} \text{ where } \mathcal{A} \in \mathcal{A}^\star.$$

The degree of tolerated inequality is captured by δ and higher δ values are associated with higher tolerance. We use exact and approximate DP to distinguish between $\delta = 0$ and $\delta > 0$.

2.5 Fair Influence Maximization

2.5.1 Cardinal Welfare Theory Background

Following the cardinal welfare theory [26], our aim is to design welfare functions to measure the goodness of the choice of influencers. Cardinal welfare theory proposes a set of principles and welfare functions that are expected to satisfy these principles. Given two utility vectors, these principles determine if they are indifferent or one of them is preferred. For ease of exposition, let W denote this welfare function defined over the utilities of all individuals in the population (we

will formalize W shortly). Then the existing principles of social welfare theory can be summarized as follows. Throughout this section, without loss of generality, we assume all utility vectors belong to $[0,1]^N$.

(1) Monotonicity. If $\boldsymbol{u} \prec \boldsymbol{u}'$, then $W(\boldsymbol{u}) < W(\boldsymbol{u}')$.[2] In other words, if $\boldsymbol{u}'$ Pareto dominates $\boldsymbol{u}$, then W should strictly prefer $\boldsymbol{u}'$ to $\boldsymbol{u}$. This principle also appears as *levelling down* objection in political philosophy [143].

(2) Symmetry. $W(\boldsymbol{u}) = W\left(P(\boldsymbol{u})\right)$, where $P(\boldsymbol{u})$ is any element-wise permutation of $\boldsymbol{u}$. According to this principle, W does not depend on the naming or labels of the individuals, but only on their utility levels.

(3) Independence of Unconcerned Individuals. Let $(\boldsymbol{u}|^c b)$ be a utility vector that is identical to $\boldsymbol{u}$, except for the utility of individual c which is replaced by a new value b. The property requires that for all $c, b, b', \boldsymbol{u}$ and $\boldsymbol{u}'$, $W(\boldsymbol{u}|^c b) < W(\boldsymbol{u}'|^c b) \Leftrightarrow W(\boldsymbol{u}|^c b') < W(\boldsymbol{u}'|^c b')$. Informally, this principle states that W should be independent of individuals whose utilities remain the same.

(4) Affine Invariance. For any $a > 0$ and b, $W(\boldsymbol{u}) < W(\boldsymbol{u}') \Leftrightarrow W(a\boldsymbol{u} + b) < W(a\boldsymbol{u}' + b)$ i.e., the relative ordering is invariant to a choice of numeraire.

($\star$5) Transfer Principle [57, 145]. Consider individuals i and j in utility vector $\boldsymbol{u}$ such that $u_i < u_j$. Let $\boldsymbol{u}'$ be another utility vector that is identical to $\boldsymbol{u}$ in all elements except i and j where $u'_i = u_i + \delta$ and $u'_j = u_j - \delta$ for some $\delta \in (0, (u_j - u_i)/2)$. Then, $W(\boldsymbol{u}) < W(\boldsymbol{u}')$. Informally, transferring utility from a high-utility to a low-utility individual should increase social welfare.

It is well-known that any welfare function W that satisfies the first four principles is additive and in the form of $W_\alpha(\boldsymbol{u}) = \Sigma_{i=1}^N u_i^\alpha/\alpha$ for $\alpha \neq 0$ and $W_\alpha(\boldsymbol{u}) = \Sigma_{i=1}^N \log(u_i)$ for $\alpha = 0$. Further, for $\alpha < 1$ the last principle is also satisfied. In this case α can be interpreted as an inequality aversion parameter, where smaller α values exhibit more aversion towards inequalities. We empirically investigate the effect of α in Section 2.6.

[2] $\prec$ means $\boldsymbol{u}_c \leq \boldsymbol{u}'_c$ for all $c \in \mathcal{C}$ and $\boldsymbol{u}_c < \boldsymbol{u}'_c$ for some $c \in \mathcal{C}$.

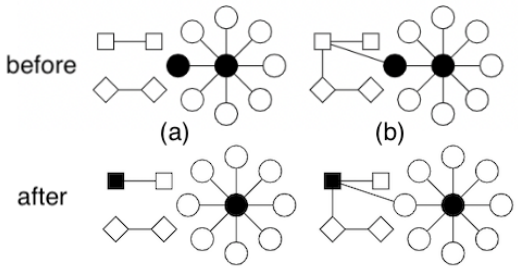


Figure 2.1: The effect of network structure and in particular between-community edges on coupling of the utilities of communities. The figure shows two sample networks consisting of three communities, differentiated by shape: (a) is the same as (b) except that between-community edges are removed. Black fillings show the choice of influencers. We further assume p is small enough such that influence spread dissipates after one step. Transferring an influencer from circles to squares (top to bottom panel) affects the utility of diamonds in (b) but not in (a).

2.5.2 Group Fairness and New Principles

Applying the cardinal social welfare framework to influence maximization problems comes with new challenges. We next highlight these challenges and demonstrate our approach.

First, the original framework of cardinal welfare theory defines the welfare function over individuals. This is equivalent to seeking equality in the probability that each individual will be influenced, similar to the work of [72]. It is notoriously hard to achieve individual fairness in practice, e.g., in [63] the authors explore this in the machine learning context. The problem is further exacerbated in influence maximization because it is not always possible to spread the influence to isolated or poorly connected individuals effectively. Therefore, we focus on group fairness whereby the utility of each individual is defined as the average utility of the members of that community. Let u_c denote the average utility of community c. With this group-wise view, a welfare function can be written in terms of the average utilities over communities e.g., $W_\alpha(\boldsymbol{u}) = \Sigma_{i=1}^{N} u_i^\alpha/\alpha = \Sigma_{c\in\mathcal{C}} N_c u_c^\alpha/\alpha$.

Moreover, while principles 1-4 can be easily extended to our influence maximization problem, this is not the case for the transfer principle. More precisely, in the influence maximization problem it might not be feasible to directly transfer utilities from one community to another

without affecting the utilities of other communities. We highlight this effect with an example, see Figure 2.1. In this figure, each community is represented by a distinct shape. The two networks (a) and (b) are identical except that between community edges are removed in network (a) (i.e., disconnected communities). The solid black vertices determine the choice of influencers. In network (b), if we transfer an influencer vertex from *circles* to *squares* according to Figure 2.1 (top to bottom panel), it will indirectly affect *diamonds* as well. This effect is absent in network (a) as there are no between-community edges to allow the spread of influence across communities. In network (a), the transfer principle prefers the resulting utility vector after the transfer. However, this principle cannot be applied to network (b) as the utilities of more than one community is modified after the transfer. Additionally, even when direct transfer is possible, it can be the case that there is no symmetry in the amount of utility gained by low-utility community and the amount of utility lost by high-utility community after the transfer. To address both of these shortcomings we introduce the *influence transfer principle* as a generalization of the transfer principle for influence maximization problems.

Similar to the original transfer principle, we consider solutions in which influencer vertices are transferred from one community to another community. Without loss of generality, we focus on the case where only one influencer vertex is transferred between the two communities. We refer to such solutions as neighboring solutions. Clearly, transfer of more than one influencer vertex can be seen as a sequence of transfers between neighboring solutions.

(5) Influence Transfer Principle. Let $\mathcal{A}$ and $\mathcal{A}' \in \mathcal{A}^\star$ be two neighboring solutions with corresponding utility vectors $\boldsymbol{u} = \boldsymbol{u}(\mathcal{A})$ and $\boldsymbol{u}' = \boldsymbol{u}(\mathcal{A}')$. Suppose the elements of $\boldsymbol{u}$ and $\boldsymbol{u}'$ are sorted in ascending order. We also assume after the transfer, the ordering of the utilities stays the same across $\boldsymbol{u}$ and $\boldsymbol{u}'$.

If $\Sigma_{\kappa\in\mathcal{C}:\kappa\leq c} N_\kappa(u'_\kappa - u_\kappa) \geq 0 \ \forall c \in \mathcal{C}$ and $u'_c > u_c$ for some $c \in \mathcal{C}$, then $W(\boldsymbol{u}) < W(\boldsymbol{u}')$.

Informally, influence transfer principle states that in a desirable transfer of utilities, the magnitude of the improvement in lower-utility communities should be at least as high as the magnitude

of decay in higher-utility communities while enforcing that at least one low-utility community receives a higher utility after the transfer. The original transfer is a special case of influence transfer principle when communities are disconnected and utilities transferred remain the same.

Next, we study whether any of the welfare functions that satisfy the first 4 principles satisfy the influence transfer principle. In Proposition 4, we show any additive and strictly concave function satisfies influence transfer principle. Since functions that satisfy the first 4 principles are strictly concave for $\alpha < 1$, the influence transfer principle is automatically satisfied in this regime. We defer all proofs to the Appendix B.

Proposition 4. *Any strictly concave and additive function satisfies influence transfer principle.*

To measure inequality, notion of utility gap (or analogous notions such as ratio of utilities) is commonly used [72, 171]. Utility gap measures the difference between the utilities of a pair of communities. In this work, we focus on the maximum utility gap, i.e., the gap between communities with the highest and lowest utilities (utility gap henceforth). For a utility vector $\boldsymbol{u}$, we define $\Delta(\boldsymbol{u}) = \max_{c \in \mathcal{C}} u_c - \min_{c \in \mathcal{C}} u_c$ to denote the utility gap. Fair interventions are usually motivated by the large utility gap before the intervention [123]. In [72], the authors have shown that in social networks the utility gap can further increase after an algorithmic influence maximizing intervention. We extend this result to the entire class of welfare functions that we study in this work and we notice that the utility gap can increase even if we optimize for these welfare functions. This is a surprising result since, unlike the influence maximization objective, these welfare functions are designed to incorporate fairness, yet we may observe an increase in the utility gap. We now introduce another principle which aims to address this issue. Again we focus on neighboring solutions.

(6) Utility Gap Reduction. Let $\mathcal{A}$ and $\mathcal{A}' \in \mathcal{A}^\star$ be two neighboring solutions with corresponding utility vectors $\boldsymbol{u} = \boldsymbol{u}(\mathcal{A})$ and $\boldsymbol{u}' = \boldsymbol{u}(\mathcal{A}')$. If $\Sigma_{c \in \mathcal{C}} N_c u_c \leq \Sigma_{c \in \mathcal{C}} N_c u'_c$. and $\Delta(\boldsymbol{u}) > \Delta(\boldsymbol{u}')$ then $W(\boldsymbol{u}) < W(\boldsymbol{u}')$.

The utility gap reduction simply states that the welfare function should prefer the utility vector whose total utility is at least as high as the other vector and also has smaller utility gap. We now show that, in general, it is not possible to design a welfare function that obeys the utility gap reduction principle along with the other principles.

Proposition 5. *Let W be a welfare function that obeys principles 1-5. Then there exists an instance of influence maximization where W does not satisfy the utility gap reduction.*

Next, we show on a special class of networks, i.e., networks with disconnected communities, the utility gap reduction principle is satisfied in all influence maximization problems.

Proposition 6. *Let W be a welfare function that obeys principles 1-5. If the communities are disconnected, then W also satisfies the utility gap reduction principle.*

Propositions 5 and 6 and their proofs establish new challenges in fair influence maximization. These challenges arise due to the coupling of the utilities as a result of the network structure and more precisely the between-community edges. The results in Propositions 5 and 6 leave open the following question: "In what classes of networks, there exists a welfare function that satisfies all the 6 principles over all instances of influence maximization problems?" As an attempt to answer this question, we empirically show that over various real and synthetic networks including stochastic block models, there exist welfare functions that obey all of our principles. We conclude this section by the following three remarks.

Remark 1 (Application to Other Settings)**.** *Our welfare-based framework can be* theoretically *applied to different graph-based problems (e.g., facility location) but algorithmic solution is domain-dependent. The choice of influence maximization is motivated by evidence about discrimination studied in previous work [151, 171].*

Remark 2 (Relationship between Principles and Fairness)**.** *Monotonicity ensures there is no wastage of utilities. Symmetry enforces the decision-maker to* not *discriminate based on communities' names. According to the Independence of Unconcerned Individuals, between two solutions*

(choices of influencers) only those individuals/communities whose utilities change should impact the decision-maker's preference. Affine Invariance is a natural requirement that the preferences over different solutions should not change based on the choice of numeraire. Finally, the Transfer Principle promotes solutions that are more equitable.

Remark 3 (Selecting the Inequality Aversion Parameter in Practice)**.** *In our approach, α is a user-selected parameter that the user can vary to tune the trade-off between efficiency and fairness. Leaving the* single *parameter α in the hands of the user is a benefit of our approach since the user can inspect the solution as α is varied to select their preferred solution. Since a single parameter must be tuned, this can be done without the need for a tailored algorithm. In particular, we recommend that α be either selected by choosing among a moderate number of values and picking the one with the most desirable behavior for the user or by using the bisection method. Typically, choosing α will reduce to letting the user select how much utility gap they are willing to tolerate: they will select the largest possible value of α for which the utility gap is acceptable.*

2.5.3 Group Fairness and Welfare Maximization

The welfare principles reflect the preferences of a fair decision-maker between a pair of solutions. Thus a welfare function that satisfies all the principles would always rank the preferred (in terms of fairness and efficiency) solution higher. As a result, we can maximize the welfare function to get the most preferred solution.

We show that the two classes of welfare functions $W_\alpha(\boldsymbol{u}) = \Sigma_{i=1}^N u_i^\alpha/\alpha$ for $\alpha < 1, \alpha \neq 0$ and $W_\alpha(\boldsymbol{u}) = \Sigma_{i=1}^N \log(u_i)$ for $\alpha = 0$ satisfy 5 of our principles. Hence as a natural notion of fairness we can define a fair solution to be a choice of influencers with the highest welfare as defined in the following optimization problem.

$$\underset{\mathcal{A}\in\mathcal{A}^\star}{\text{maximize}} \quad W_\alpha(\boldsymbol{u}(\mathcal{A})). \tag{2.3}$$

Lemma 3. *In the influence maximization problem, any welfare function that satisfies principles 1-5 is monotone and submodular.*

It is well-known that to maximize any monotone submodular function, there exists a greedy algorithm with a $(1 - 1/e)$ approximation factor [136] which we can also use to solve the welfare maximization problem.

Each choice of the inequality aversion parameter α results in a different welfare function and hence a fairness notion. A decision-maker can directly use these welfare functions as objective of an optimization problem and study the trade-off between fairness and total utility by varying α, see Section 5.

2.5.4 Connection to Existing Notions of Fairness

Our framework allows for a spectrum of fairness notions as a function of α. It encompasses as a special case *leximin fairness*[3], a sub-class of MMF, for $\alpha \to -\infty$. Proportional fairness [43], a notion for resource allocation problems, is also closely connected to the welfare function for $\alpha = 0$.

It is natural to ask which of the fairness principles are satisfied by the existing notions of fairness for influence maximization. As we discussed in Section 2.4, the existing notions of fairness are imposed by adding constraints to the influence maximization problem. However, our welfare framework directly incorporates fairness into the objective. In order to facilitate the comparison, instead of the constrained influence maximization problems we consider an equivalent reformulation in which we bring the constraints into the objective via the characteristic function of the feasible set. We then have a single objective function which we can treat as the welfare

[3]Leximin is subclass of MMF. According to its definition, among two utility vectors, leximin prefers the one where the worst utility is higher. If the worst utilities are equal, leximin repeats this process by comparing the second worst utilities and so on.

	Mono.	Sym.	Ind. of Unconcerned	Affine	Inf. Transfer	Gap Red.
Exact DP	✗ (Prop. 5)	✓	✗ (Prop. 8)	✓	✗ (Prop. 11)	✓ (Prop. 15)
Approx. DP	✗ (Prop. 6)	✓	✗ (Prop. 8)	✗	✗ (Prop. 11)	✗ (Prop. 15)
DC	✓ (Cor. 1)	✗	✗ (Prop. 9)	✗	✗ (Prop. 12)	✗ (Prop. 16)
MMF	✓ (Cor. 1)	✓	✗ (Prop. 10)	✓	✗ (Prop. 13)	✗ (Prop. 17)
Utilitarian	✓ (Cor. 1)	✓	✓	✓	✗ (Prop. 14)	✗ (Prop. 18)
Welfare	✓	✓	✓	✓	✓	✗ (Prop. 2)

Table 2.1: Summary of the properties of different fairness notions through the lens of welfare principles for influence maximization.

function corresponding to the fairness constrained problem. More formally, given an influence maximization problem and fairness constraints written as a feasible set $\mathcal{F}$

$$\max_{\mathcal{A}\in\mathcal{A}^\star} \sum_{c\in\mathcal{C}} N_c u_c(\mathcal{A}) \quad \text{s.t.} \quad \boldsymbol{u}(\mathcal{A}) \in \mathcal{F}.$$

We consider the following equivalent optimization problem

$$\max_{\mathcal{A}\in\mathcal{A}^\star} \sum_{c\in\mathcal{C}} N_c u_c(\mathcal{A}) + \mathcal{I}_{\mathcal{F}}(\boldsymbol{u}(\mathcal{A})) := \max_{\mathcal{A}\in\mathcal{A}^\star} W_{\mathcal{F}}(\boldsymbol{u}(\mathcal{A})),$$

in which $\mathcal{I}_{\mathcal{F}}(\boldsymbol{u})$ is equal to 0 if $\boldsymbol{u} \in \mathcal{F}$ and $-\infty$ otherwise. Using this new formulation, we can now examine each of the existing notions of fairness though the lens of the welfare principles. Given the new interpretation, to show that a fairness notion does not satisfy a specific principle, it suffices to show there exist solutions $\mathcal{A}, \mathcal{A}' \in \mathcal{A}^\star$ and corresponding utility vectors $\boldsymbol{u} = \boldsymbol{u}(\mathcal{A})$ and $\boldsymbol{u}' = \boldsymbol{u}(\mathcal{A}')$ such that the principle prefers $\boldsymbol{u}$ over $\boldsymbol{u}'$ but $W_{\mathcal{F}}(\boldsymbol{u}) < W_{\mathcal{F}}(\boldsymbol{u}')$. The results are summarized in Table 2.1 where in addition to comparing with the previous notions introduced in Section 2.4, we compare with the utilitarian notion i.e., Problem (2.1). We provide formal proofs for each entry of Table 2.1 in Appendix B.

We observe that none of the previously defined notions of fairness for influence maximization satisfies all of our principles and each existing notion violates at least 3 out of 6 principles. We point out that exact DP is the only notion that satisfies the utility gap reduction. However, this

comes at a cost as enforcing exact DP may result in significant reduction in total utility in the fair solution compared to the optimal unconstrained solution [54].

We evaluate our approach in terms of both the total utility or spread of influence (to account for efficiency) and utility gap (to account for fairness). We show by changing the inequality aversion parameter, we can effectively trade-off efficiency with fairness. As baselines, we compare with DC and MMF. To the best our knowledge, there is no prior work that handles DP constraints over the utilities. We follow the approach of [175] for both problems and view these problems as a multi-objective submodular optimization with utility of each each community being a separate objective. They propose an algorithm and implementation with asymptotic $(1-1/e)$ approximation guarantee which we also utilize here. We use *Price of Fairness* (PoF), defined as the percentage loss in the total influence spread as a measure of efficiency. Precisely, PoF $:= 1 - \mathrm{OPT}^{\mathrm{fair}}/\mathrm{OPT}^{\mathrm{IM}}$ in which $\mathrm{OPT}^{\mathrm{fair}}$ and $\mathrm{OPT}^{\mathrm{IM}}$ are the the total influence spread, with and without fairness. Hence PoF$\in [0, 1]$ and smaller values are more desirable. The normalization in PoF allows for a meaningful comparison between networks with different sizes and budgets as well as between different notions of fairness. In the PoF calculations, we utilize the generic greedy algorithm [105] to compute $\mathrm{OPT}^{\mathrm{IM}}$. To account for fairness, we compare the solutions in terms of the utility gap. Analogous measures are widely used in fairness literature [85] and more recently in graph-based problems [72, 171]. We also note that our framework ranks solutions based on their welfare and does not directly optimize utility gap, as such our evaluation metric of fairness does not favor any particular approach.

We perform experiments on both synthetic and real networks. We study two applications: community preparedness against landslide incidents and suicide prevention among homeless youth. We discuss the latter in Appendix B. In the synthetic experiments, we use the *stochastic block model* (SBM) networks, a widely used model for networks with community structure [70]. In SBM networks, vertices are partitioned into disjoint communities. Within each community c, an edge between two vertices is present independently with probability q_c. Between any two vertices in

communities c and c', an edge exists independently with probability $q_{cc'}$ and typically $q_c > q_{cc'}$ to capture homophily [125]. SBM captures the community structure of social networks [81]. We report the average results over 20 random instances and set $p = 0.25$ in all experiments.

Landslide Risk Management in Sitka, Alaska. Sitka, Alaska is subject to frequent landslide incidents. In order to improve communities' preparedness, an effective approach is to instruct people on how to protect themselves before and during landslide incidents. Sitka has a population of more than 8000 and instructing everyone is not feasible. Our goal is to select a limited set of individuals as peer-leaders to spread information to the rest of the city. The Sitka population is diverse including different age groups, political views, seasonal and stable residents where each person can belong to multiple groups. These groups differ in their degree of connectedness. This makes it harder for some groups to receive the intended information and also impacts the cost of imposing fairness.

Since collecting the social network data for the entire city is cumbersome, we assume a SBM network and use in-person semi-structured interview data from 2018-2020 with members of Sitka to estimate the SBM parameters. Using the interview responses in conjunction with the voter lists, we identified 5940 individuals belonging to 16 distinct communities based on the intersection of age groups, political views, arrival time to Sitka (to distinguish between stable and transient individuals). The size of the communities range from 112 (stable, democrat and 65+ years of age) to 693 (republican, transient fishing community, age 30-65). See Appendix B for details on the estimation of network parameters.

2.6 Computational Results

Figure 2.2 summarizes results across different budget values K ranging from 2% to 10% of the network size N for our framework (different α values) as well as the baselines. In the left panel, we observe that as α decreases, our welfare-based framework further reduces the utility gap,

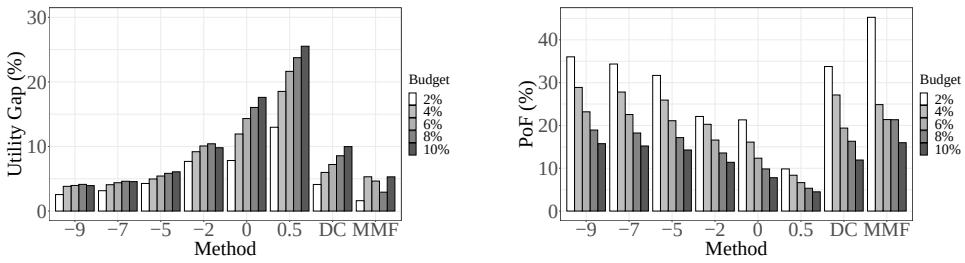


Figure 2.2: Left and right panels: utility gap and PoF for different K and α values for our framework and baselines.

achieving lower gap than DC and competitive gap as MMF. As we noted in Section 2.5.4, our framework recovers leximin (which has stronger guarantees than MMF) as $\alpha \to -\infty$, though we show experimentally that this is achieved with moderate values of α. Overall, utility gap shows an increasing trend with budget, however the sensitivity to budget decreases when more strict fairness requirements are in place, e.g. in MMF and $\alpha = -9.0$. From the right panel, PoF varies significantly across different approaches and budget values surpassing 40% for MMF. This is due to the stringent requirement of MMF to raise the utility of the worst-off as much as possible. Same holds true for lower values of α as they exhibit higher aversion to inequality. The results also indicate that PoF decreases as K grows which captures the intuition that fairness becomes less costly when resources are in greater supply. Resource scarcity is true in many practical applications, including the landslide risk management domain which makes it crucial for decision-makers to be able to study different fairness-efficiency trade-offs to come up with the most effective

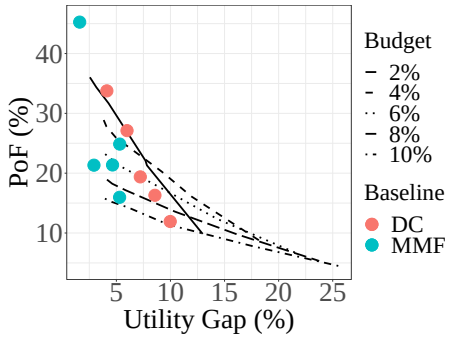


Figure 2.3: PoF vs. utility gap trade-off curves. Each line corresponds to a different budget K across different α values.

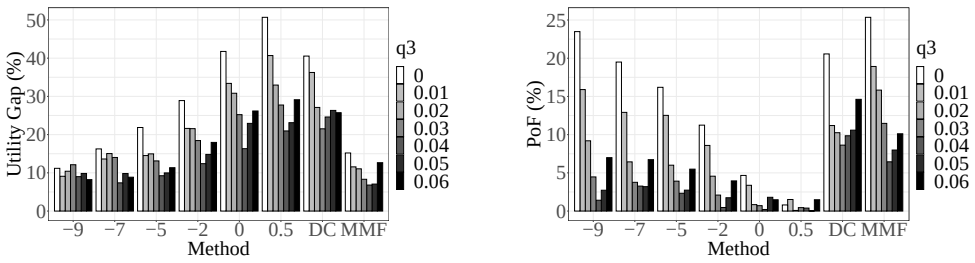


Figure 2.4: Utility gap and PoF for various levels q_3. All results are compared across different values of α and the baselines.

plan. Figure 2.3 depicts such trade-off curves where each line corresponds to a different budget value across the range of α. Previous work only allows a decision-maker to choose among a very limited set of fairness notions regardless of the application requirements. Here, we show that our framework allows one to choose α to meaningfully study the PoF-utility gap trade-offs. For example, given a fixed budget and a tolerance on utility gap, one can choose an α with the lowest PoF. We now investigate the effect of relative connectedness. We provide the effect of relative community size in Appendix A.

Relative Connectedness. We sample SBM networks consisting of 3 communities each of size 100 where communities differ in their degree of connectedness. We set $q_1 = 0.06, q_2 = 0.03, q_3 = 0.0$ to obtain three communities with high, mid and low relative connectedness. We choose these values to reflect asymmetry in the structure of different communities which mirrors real world scenarios since not every community is equally connected. We set between-community edge probabilities $q_{cc'}$ to 0.005 for all c and c' and $K = 0.1N$. We gradually increase q_3 from 0.0 to 0.06. Results are summarized in Figure 2.4, where each group of bars correspond to a different approach. We observe as q_3 increases utility gap and PoF decrease until they reach a minimum around at around $q_3 = 0.03$. From this point, the trend reverses. This U-shaped behavior is due to structural changes in network. More precisely, for $q_3 < 0.03$ we are in the high-mid-low connectedness regime for the three groups, where the third community receives the minimum utility. As a result, as q_3 increases it becomes more favorable to choose more influencer vertices in

this community which in turn reduces the utility gap. For $q_3 > 0.03$, the second community will be become the new worst-off community due its lowest connectedness. Hence, further increase in q_3 causes more separation in connectedness and we see previous behavior in reverse. Thus, by further increasing q_3, communities 1 and 3 receive more and more influencer vertices. This behavior translates to PoF as the relative connectedness of communities impacts how *hard* it is to achieve a desired level of utility gap. Finally, we see that the U-shaped behavior is skewed, i.e., we observe higher gap and PoF in lower range of q_3 which is due to higher gap in connectedness of communities. We can also compare the effect of relative connectedness and community size (see Appendix B).

We observe that connectedness has a more significant impact on PoF (up to 25%) compared to community size (less than 4%). In other words, when communities are structurally different it is more costly to impose fairness. This is an insightful result given that in different applications we may encounter different populations with structurally different networks. Utility gap on the other hand is affected by both size and connectedness. Finally while our theory indicates that in the network setting, no welfare function can satisfy all principles including utility gap reduction over all instances of the influence maximization, we observe that our class of welfare functions satisfies all of the desiderata on the class of networks that we empirically study. Our theoretical results showed this for a special case of networks with disconnected communities. In particular, we see higher PoF is accompanied by lower utility gap which complies with utility gap reduction principle.

2.7 Conclusion and Broader Impact

As the empirical evidence highlighting ethical side effects of algorithmic decision-making is growing [8, 128], the nascent field of algorithmic fairness has also witnessed a significant growth. It is

well-established by this point that there is no universally agreed-upon notion of fairness, as fairness concerns vary from one domain to another [135, 27]. The need for different fairness notions can also be explained by theoretical studies that show that different fairness definitions are often in conflict with each other [110, 51, 74]. To this end, most of the literature on algorithmic fairness proposes different fairness notions motivated by different ethical concerns. A major drawback of this approach is the difficulty of comparing these methods against each other in a systematic manner to choose an appropriate notion for the domain of interest. Instead of following this trend, we propose a unifying framework controlled by a single parameter that can be used by a decision-maker to systematically compare different fairness measures which typically result in different (and possibly also problem-dependent) trade-offs. Our framework also accounts for the social network structure while designing fairness notions – a consideration that is mainly overlooked in the past. Given these two contributions, it is perceivable that our approach can be used in many of the public health interventions such as suicide, HIV or Tuberculous prevention that rely on social networks. This way, the decisions-makers can compare a menu of fairness-utility trade-offs proposed by our approach and decide which one of these trade-offs are more desirable without a need to understand the underlying mathematical details that are used in deriving these trade-offs.

There are crucial considerations when deploying our system in practice. First, cardinal welfare is one particular way of formalizing fairness considerations. This by no means implies that other approaches for fairness e.g. equality enforcing interventions should be completely ignored. Second, we have assumed that the decision-maker has the full knowledge of the network as well as possibly protected attributes of the individuals which can be used to define communities. Third, while our experimental evaluation is based on utilizing a greedy algorithm, it is conceivable that this greedy approximation can create complications by imposing undesirable biases that we have not accounted for. Intuitively (and as we have seen in our experiments) the extreme of inequality

aversion ($\alpha \to -\infty$) can be used as a proxy for pure equality. However, the last two concerns require more care and we leave the study of such questions as future work.

Part II

Algorithmic Fairness under Observational Data

Chapter 3

Fair and Efficient Housing Allocation Policy Design

3.1 Introduction

We study the problem of designing policies to effectively match heterogeneous individuals to scarce resources of different types. We consider the case where both individuals and resources arrive stochastically over time. Upon arrival, each individual is assigned to a queue where they wait to be matched to a resource. This problem arises in several public systems such as those providing social services, posing unique challenges at the intersection of efficiency and fairness. In particular, the joint characteristics of individuals and their matched resources determine the effectiveness of an allocation policy, making it crucial to match individuals with the right type of resource. Furthermore, when a resource becomes available, a decision-maker should decide whom among the individuals waiting in various queues should receive the resource which impacts the wait time of different individuals. In addition, since there are insufficient resources to meet demand, there are inherent fairness considerations for designing such policies.

We are particularly motivated by the problem of allocating housing resources among individuals experiencing homelessness. According to the U.S. Department of Housing and Urban Development (HUD), more than 580,000 people experience homelessness on a given night [87]. The Voices of Youth Count study found youth homelessness has reached a concerning prevalence level in the United States; one in 30 teens (13 to 17) and one in 10 young adults (18 to 25) experience at least one night of homelessness within a 12-month period, amounting to 4.2 million persons a year [133]. Housing interventions are widely considered as the key solution to address homelessness [139]. In the U.S., the government funds programs that assist homeless using different forms of housing interventions and services [177]. The HMIS database collects information on the provision of these services.

Unfortunately, the number of homeless individuals in the U.S. far exceeds the available resources which necessitates strategic allocation to maximize the intervention's effectiveness. Many

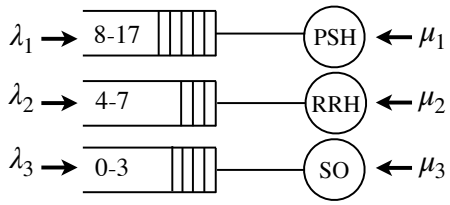


Figure 3.1: NST-recommended resource allocation policy utilized by housing allocation agencies in the homelessness context. The policy is in the form of a resource eligibility structure. According to this figure, individuals with score eight and above qualify for PSH, score 4 to 7 are assigned to the RRH wait list and finally individuals who score below 4 are not assigned to any of the housing interventions.

communities have attempted to address this problem by creating coordinated community responses, typically referred to as Coordinated Entry Systems (CES). In such systems, most agencies within a community pool their housing resources in a centralized system called a Continuum of Care (CoC). A CoC is a regional or local planning body that coordinates housing and services funding—primarily from HUD—for people experiencing homelessness. Individuals in a given CoC who seek housing are first assessed for eligibility and vulnerability and those identified as having the greatest need are matched to appropriate housing resources [157]. For example, in the context of youth homelessness, the most widely adopted tool for assessing vulnerability is the Transition Age Youth-Vulnerability Index-Service Prioritization Decision Assistance Tool (TAY-VI-SPDAT): Next Step Tool (NST), which was developed by OrgCode Consulting, Corporation for Supportive Housing (CSH), Community Solutions, and Eric Rice. OrgCode claims that hundreds of CoC's in the USA, Canada and Australia have adopted this tool [140]. After assessment, each individual receives a vulnerability score ranging from 0 to 17. One of the main challenges that CoC's face is how to use the information about individuals to decide what housing assistance programs should be available to a particular homeless individual. In many communities, based on the recommendations provided in the NST tool documentation, individuals who score 8 to 17 are considered as "high risk" and are prioritized for resource-intensive housing programs or Permanent Supportive Housing (PSH). Those who score in the 4-7 range are typically assigned to short-term rental subsidy programs or Rapid-ReHousing (RRH) and those with score below 4 are eligible for services

that meet basic needs which we refer to as Service Only (SO) [158]. Figure 3.1 depicts how the individuals are matched to resources according to the status-quo policy.

The aforementioned policy can be viewed as a *resource eligibility structure* as from the onset, it determines the resources an individual is eligible for. Such policies have the advantage of being interpretable, i.e., it is easy to explain why a particular allocation is made. Earlier work shows that most communities follow the policy recommendations when assigning housing [158]. However, controversy has surrounded the use of these cut scores and as of December 2020, OrgCode has called for new approaches to using the data collected by HMIS [141]. There is also an overwhelming desire on the part of HUD to design systematic and data-driven housing policies, including the design of the cut scores and the queues that they induce [177]. Currently, the cut scores are not tied to the treatment effect of interventions or the relative arrival rate of individuals and resources in the respective queues. This is problematic as it is not evidently clear that assigning high-scoring and mid-scoring individuals to particular housing interventions, such as PSH or RRH, actually increases their chances of becoming stably housed. Additionally, there may not be enough resources to satisfy the needs of all individuals matched to a particular resource, resulting in long wait times. Prolonged homelessness may in turn increase the chances of exposure to violence, substance use, etc., or individuals dropping out of the system.

In particular, OrgCode and others have called for a new equity focus to how vulnerability tools are linked to housing allocation [141, 127]. Despite recent efforts to understand and mitigate disparities in homelessness, current system suffers from a significant gap in the prevalence of homelessness across different groups. For example, studies show that most racial minority groups experience homelessness at higher rates than Whites [78]. Also, recent work has revealed that PSH outcomes are worse for Black clients in Los Angeles [127] and based on the same HMIS data used in present study, Black, Latinx, and LGBQ youth have been shown to experience worse housing outcomes [89]. Addressing these disparities requires an understanding of the distribution

of the individuals vulnerability to homelessness, the heterogeneity in the treatment affect and the associations with protected attributes such as race, gender, or age.

In this work, we build on the literature on causal inference and queuing theory and we propose a methodology to use historical data about the waitlisted individuals and their allocated resources to evaluate and optimize new resource allocation policies that take policy effectiveness, fairness and wait time into account. We make the following contributions:

- We model the policy optimization problem as a multi-class multi-server queuing system between heterogeneous individuals and resources that arrive over time. We extend the literature on queuing theory by proposing a data-driven methodology to construct the model from observational data. Specifically, we use tools from modern causal inference to learn the treatment effect of the interventions from data and construct the queues by grouping individuals that have similar average treatment effects.

- We propose interpretable policies that take the form of a resource eligibility structure, encoding the resource types that serve each queue. We provide an MIO formulation to optimize the eligibility structure that incorporates flexibly defined fairness considerations or other linear domain-specific constraints. The MIO maximizes the policy effectiveness and guarantees minimum wait time.

- Using HMIS data, we conduct a case study to demonstrate the effectiveness of our approach. Our results indicate superior performance along policy effectiveness, fairness and wait time. Precisely, we are able to obtain wait time as low as a fully FCFS policy while improving the rate of exit from homelessness for traditionally underserved or vulnerable groups (7% for the Black individuals and 15% higher for youth below 17 years old) and overall.

The remainder of this chapter is organized as follows. In Section 3.2, we review the related literature. In Section 3.3, we introduce the policy optimization problem. In Section 3.4, we propose our data-driven methodology for solving the policy optimization problem. Finally, we summarize

our numerical experiments and present a case study using HMIS data on youth experiencing homelessness in Section 3.5. Proofs and detailed numerical results are provided in the Appendix C.

3.2 Related Work

This work is related to several streams of literature which we review. Specifically, we cover queuing theory as the basis of our modelling framework. We also position our methodology within the literature on data-driven policy optimization and causal inference. We conclude by highlighting recent works on fairness in resource allocation.

A large number of scarce resource allocation problems give rise to one-sided queuing models. In these models, resources are allocated upon arrival, whereas individuals queue before being matched. Examples are organ matching [14] and public housing assignment [103]. One stream of literature studies dynamic matching policies to find asymptotically optimal scheduling policies under conventional heavy traffic conditions [10, 121]. Another stream focuses on the system behavior under FCFS service discipline aiming to identify conditions that ensure the stability of the queuing system and characterize the steady-state matching flow rates, i.e., the average rate of individuals of a given queue (or customer class) that are served by a particular resource (server) [45, 67]. These works only focus on minimizing delay and do not explicitly model the heterogeneous service value among the customers. Recently, [60] studied one-sided queuing system where resources are allocated to the customer with the highest score (or index), which is the sum of the customer's waiting score and matching score. The authors derive a closed-form index that optimizes the steady-state performance subject to specific fairness considerations. Their proposed fairness metric measures the variance in the likelihood of getting service before abandoning the queue. Contrary to their model, we consider FCFS policies subject to resource eligibility structures which we optimize over. Our model is based on the policies currently being implemented for housing allocation among homeless individuals and are interpretable by design. In addition, our

model allows for a more general class of fairness constrained commonly used in practice including fairness in *allocation* and *outcome*.

Our approach builds upon [3], in which the authors study the problem of designing a matching topology between customer classes and servers under a FCFS service discipline. They focus on finding matching topologies that minimize the customers' waiting time and maximize matching rewards obtained by pairing customers and servers. The authors characterize the average steady-state wait time across all customer classes in terms of the structure of the matching model, under heavy-traffic condition. They propose a quadratic program (QP) to compute the steady-state matching flows between customers and servers and prove the conditions under which the approximation is exact. We build on the theoretical results in [3] to design resource eligibility structures that match heterogeneous individuals and resources in the homelessness setting. Contrary to the model in [3], we do not assume that the queues or the matching rewards are given a priori. Instead, we propose to use observational data from historical policy to learn an appropriate grouping of individuals into distinct queues, estimate the matching rewards, and evaluate the resulting policies.

Another stream of literature focuses on designing data-driven policies, where fairness considerations have also received significant attention due to implicit or explicit biases that models or the data may exhibit [34, 59, 148, 106]. In [34], the authors propose a data-driven model for learning scoring policies for kidney allocation that matches organs at their time of procurement to available patients. Their approach satisfies linear fairness constraints approximately and does not provide any guarantees for wait time. In addition, they take as input a model for the matching rewards (i.e., life years from transplant)to optimize the scoring policy. In [13], the authors propose a data-driven mixed integer program with linear fairness constraints to solve a similar resource allocation which provides an exact, rather than an approximate, formulation. They also give an approximate solution to achieve faster run-time. We consider a class policies in the form of matching topologies that is different from scoring rules and is more closely related to the policies

implemented in practice. Such policies offer more interpretability as individuals know what resources they are eligible for from the onset. Several works have considered interpretable functional forms in policy design. For example, in [31, 98], the authors consider decision trees and develop techniques to obtain optimal trees from observational data. Their approach is purely data-driven and do not allow for explicit modelling of the arrival of resources, individuals which impact wait time. In the homelessness setting, our work is closely related to [114] which proposes a resource allocation mechanism to match homeless households to resources based on the probability of system re-entry. In this work, the authors provide a static formulation of the problem which requires frequent re-optimization and does not take the waiting time into account. In [137], the authors propose a fairness criterion that prioritizes those who benefit the most from a resource, as opposed to those who are the neediest and study the price of fairness under different fairness definitions. Similar to [114], their formulation is static and does not yield a policy to allocate resources in dynamic environments.

3.3 Housing Allocation as a Queuing System

3.3.1 Preliminaries

We model the resource allocation system as an infinite stream of heterogeneous individuals and resources that arrive over time. Each individual is characterized by a (random) feature vector $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^n$ and receives an intervention R from a finite set of treatments indexed in the set $\mathcal{R}$. We note that $\mathcal{R}$ may include "no intervention" or minimal interventions such as SO in the housing allocation setting. Using the potential outcomes framework [162], each individual has a vector of potential outcomes $Y(r) \in \mathcal{Y} \subseteq \mathbb{R}\ \forall r \in \mathcal{R}$, where $Y(r)$ is an individual's outcome when matched to resource r.

We assume having access to N historical observations $\mathcal{D} := \{(\boldsymbol{X}_i, R_i, Y_i)\}_{i=1}^N$, generated by the deployed policy, where $\boldsymbol{X}_i \in \mathcal{X}$ denotes the feature vector of the ith observation, $R_i \in \mathcal{R}$ is

the resource assigned to it and $Y_i = Y_i(R_i)$ is the observed outcome, i.e., the outcome under the resource received. A (stochastic) policy $\pi(r|\boldsymbol{x}) : \mathcal{X} \times \mathcal{R} \to [0, 1]$ maps features $\boldsymbol{x}$ to the probability of receiving resource r. We define the value of a policy as the expected outcome when the policy is implemented, i.e., $V(\pi) := \mathbb{E}[\sum_{r \in \mathcal{R}} \pi(r|\boldsymbol{X})Y(r)]$. A major challenge in evaluating and optimizing policies is that we cannot observe the counterfactual outcomes $Y_i(r), r \in \mathcal{R}, r \neq R_i$ of resources that were not received by data point i. Hence, we need to make further assumptions to identify policy values from historical data. In Section 3.4, we elaborate on these assumptions and propose our methodology for evaluating and optimizing policies from data.

We model the system as a multi-class multi-server (MCMS) queuing system where a set of resources $\mathcal{R}$ serve a finite set of individual queues indexed in the set $\mathcal{Q}$. Upon arrival, individuals are assigned to different queues based on their feature vector. For example, in the housing allocation setting and according to the recommended policy the assignment is based on the vulnerability score. We use $p : \mathcal{X} \to \mathcal{Q}$ to denote the function that maps the feature vector to a queue that the individual will join. We refer to p as the *partitioning function* (as it partitions $\mathcal{X}$ and assigns each subset to a queue) and note that it is unknown a priori. In this work, we consider partitioning functions in the form of a binary trees similar to classification trees, due to their interpretability [13]. We assume that individuals arrive according to independent Poisson processes and that inter-arrival time of resources follows an exponential distribution. These are common assumptions in queuing theory for modeling arrivals. We use $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_{|\mathcal{Q}|})$ and $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_{|\mathcal{R}|})$ to denote the vector of arrival rates of individuals and resources, respectively. We define $\lambda_{\mathcal{Q}} := \sum_{q \in \mathcal{Q}} \lambda_q$ and $\mu_{\mathcal{R}} := \sum_{r \in \mathcal{R}} \mu_r$ as the cumulative arrival rates of individuals and resources, respectively. Without loss of generality, we assume that $\lambda_q > 0 \ \forall q \in \mathcal{Q}$ and $\mu_r > 0 \ \forall r \in \mathcal{R}$.

3.3.2 Matching Policy

Once a new resource becomes available, it is allocated according to a resource eligibility structure that determines what queues are served by any particular resource. The resource eligibility structure can be equivalently represented as a matching topology $\boldsymbol{M} := [M_{qr}] \in \{0,1\}^{|\mathcal{Q}|\times|\mathcal{R}|}$, where $M_{qr} = 1$ indicates that individuals in queue q is eligible for resource r. Resources are assigned to queues in an FCFS fashion subject to matching topology $\boldsymbol{M}$. For a partitioning function p and matching topology $\boldsymbol{M}$, we denote the allocation policy by $\pi_{p,\boldsymbol{M}}(r|\boldsymbol{x})$. We concern ourselves with the long-term steady state of the system. Proposition 7 gives the necessary and sufficient conditions to arrive at a steady-state.

Proposition 7 ([2], Theorem 2.1)**.** *Given the MCMS system defined through* $(\mathcal{Q},\mathcal{R},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{M})$*, under the FCFS service discipline matching* $\boldsymbol{M}$ *admits a steady state if and only if the following condition is satisfied:*

$$\mu_{\mathscr{R}} - \sum_{r\in\mathscr{R}} \sum_{q\in\mathcal{Q}_{\mathscr{R}}(\boldsymbol{M})} \lambda_q > 0 \quad \forall \mathscr{R} \subseteq \mathcal{R}.$$

The left-hand side is the cumulative arrival rate of resources in $\mathscr{R}$ *in excess of the cumulative arrival rate of all the queues in* $\mathcal{Q}_{\mathscr{R}}$*, where* $\mathcal{Q}_{\mathscr{R}}$ *is the set of queues that are only eligible for resources in* $\mathscr{R}$*, i.e.,* $\mathcal{Q}_{\mathscr{R}} = \{q \in \mathcal{Q} : \sum_{r\in\mathcal{R}\setminus\mathscr{R}} M_{qr} = 0\}$.

According to the above result, we can define the set of *admissible matching topologies* as those that satisfy the inequality in Proposition 7. In the housing allocation problem, we assume that SO resources are abundant, i.e., $\mu_{\mathcal{R}} - \lambda_{\mathcal{Q}} > 0$. As a result, there exists at least one admissible matching: the fully connected matching topology $M_{qr} = 1\ \forall q \in \mathcal{Q}, r \in \mathcal{R}$. The abundance assumption is necessary in order to avoid overloaded queues. However, in practice housing resources are strictly preferred. As a result, we propose to study the system under the so-called *heavy traffic* regime, where the system is loaded very close to its capacity and we assume that the system utilization parameter $\rho := \mu_{\mathcal{R}}/\lambda_{\mathcal{Q}}$ approaches 1, i.e., $\rho \to 1$. In general, we assume

that $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are such that $\lambda_{\mathcal{Q}} = \rho\mu_{\mathcal{R}}$. This assumption will additionally make the analytical study of the matching system more tractable. In particular, in [3], the authors propose a quadratic program to approximate the exact steady-state flows of the stochastic FCFS matching system under heavy traffic conditions. They enforce the steady-state flows in an optimization model to find the optimal matching topology using KKT optimality conditions. We adopt the same set of constraints in the present work. We discuss in more detail when we present the final optimization formulation. We let $\boldsymbol{F} := [F_{qr}] \in \mathbb{R}_+^{|\mathcal{Q}|\times|\mathcal{R}|}$ denote the steady-state flow, where $M_{qr} = 0 \Rightarrow F_{qr} = 0$. Given a partitioning function p, the policy associated with a matching topology $\boldsymbol{M}$ is equal to $\pi_{p,\boldsymbol{M}}(r|\boldsymbol{x}) = F_{qr}/\sum_{r\in\mathcal{R}} F_{qr} = F_{qr}/\lambda_q$, in which $q = p(\boldsymbol{x})$ and the second inequality follows from the flow balance constraints. In Proposition 8 we show how the policy value can be written using the matching model parameters and treatment effect of different interventions. We define the conditional average treatment effect (CATE) of resource r and queue q as $\tau_{qr} := \mathbb{E}[Y(r) - Y(1)|P(\boldsymbol{X}) = q]\ \forall r \in \mathcal{R}, q \in \mathcal{Q}$, in which $r = 1$ is the baseline intervention. In many applications, the baseline intervention corresponds to "no-intervention" (also referred to as the control group). In the housing allocation context, we set $r = 1$ to be the SO intervention.

Proposition 8. *Given a partitioning function p, an MCMS model $(\mathcal{Q}, \mathcal{R}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{M})$, and the steady-state FCFS flow $\boldsymbol{F}$ under FCFS discipline, the value of the induced policy $\pi_{p,\boldsymbol{M}}$ is equal to:*

$$V(\pi_{p,\boldsymbol{M}}) = \frac{1}{\lambda_{\mathcal{Q}}} \sum_{q\in\mathcal{Q}} \sum_{r\in\mathcal{R}} F_{qr}\tau_{qr} + C,$$

where C is a constant that depends on the expected outcome under the baseline intervention.

3.3.3 Policy Optimization

We now introduce the policy optimization problem under the assumption that the joint distribution of $\boldsymbol{X}, Y(r), r \in \mathcal{R}$ as well as the partitioning function p is known. In Section 3.4, we propose a

methodology to construct p. Furthermore, we describe how we can use historical data to optimize new policies.

$$\mathcal{P}(p) := \max_{\boldsymbol{M} \in \mathcal{M}} V(\pi_{p,\boldsymbol{M}}). \tag{3.1}$$

In the above formulation, $\mathcal{M}$ is the set of feasible matching topologies. In addition to the constraints that impose steady-state flow, we incorporate fairness and wait time constraints in the set $\mathcal{M}$ which we describe next.

Fairness In this work, we focus on group-based notions of fairness which have been widely studied in recent years in various data-driven decision making settings [151, 13, 137, 34]. Formally, we let G be a random variable describing the group that an individual belongs to, taking values in $\mathcal{G}$. For example, G can correspond to protected features such as race, gender or age. It is also possible to define fairness with respect to other features, such as vulnerability score in the housing allocation setting. We give several examples to which our framework applies.

Example 1 (Maximin Fair in Allocation)**.** *Motivated by Rawls theory of social justice [156], maximin fairness aims to help the worst-off group as much as possible. Formally, the fairness constraints can be written as*

$$\sum_{q \in \mathcal{Q}_g} F_{qr} \geq w \quad \forall g \in \mathcal{G}, r \in \mathcal{R},$$

where w is the minimum acceptable flow across groups and $\mathcal{Q}_g \subseteq \mathcal{Q}$ is a subset of queues whose individuals belong to $G = g$. If queues contain individuals with different values of g, one should separate them by creating multiple queues with unique g. By increasing the parameter w, one is imposing more strict fairness requirements. This parameter can be used to control the trade-off between fairness and policy value. It can also be set to the highest value for which the constraint is feasible.

Example 2 (Group-based Parity in Allocation)**.** *Parity-based fairness notions strive for equal outcomes across groups.*

$$\left| \sum_{q \in \mathcal{Q}_g} F_{qr} - \sum_{q \in \mathcal{Q}_{g'}} F_{qr} \right| \leq \epsilon \quad \forall g, g' \in \mathcal{G}, r \in \mathcal{R}.$$

In words, for every resource the difference between the cumulative flow between any pair of groups should be at most ϵ, where ϵ can be used to control the trade-off between fairness and policy value.

Example 3 (Maximin Fair in Outcome)**.** *For every group, the policy value should be at least w.*

$$\frac{1}{\lambda_{\mathcal{Q}_g}} \sum_{q \in \mathcal{Q}_g} \sum_{r \in \mathcal{R}} F_{qr} \tau_{qr} \geq w \quad \forall g \in \mathcal{G}.$$

Example 4 (Group-based Parity in Outcome)**.** *The difference between the policy value for any pair of groups is at most ϵ.*

$$\left| \frac{1}{\lambda_{\mathcal{Q}_g}} \sum_{q \in \mathcal{Q}_g} \sum_{r \in \mathcal{R}} F_{qr} \tau_{qr} - \frac{1}{\lambda_{\mathcal{Q}_{g'}}} \sum_{q \in \mathcal{Q}'_g} \sum_{r \in \mathcal{R}} F_{qr} \tau_{qr} \right| \leq \epsilon \quad \forall g, g' \in \mathcal{G}.$$

In the experiments, we focus on fairness in outcome due to treatment effect heterogeneity. In other words, it is important to match individuals with the right type of resource, rather than ensuring all groups have the same chance of receiving any particular resource. Further, we adopt maximin fairness which guarantees Pareto optimal policies [148].

Wait Time Average wait time is dependent on the structure of the matching topology. For example, minimum average wait time is attainable in a fully FCFS policy where $M_{qr} = 1 \; \forall q \in \mathcal{Q}, r \in \mathcal{R}$. In [3], the authors characterize the general structural properties that impact average wait time. In particular, they show that under the heavy traffic condition, a matching system can be partitioned into a collection of complete resource pooling (CRP) subsystems that operate "almost" independently of each other. A key property of this partitioning is that individuals

that belong to the same CRP component experience the same average steady-state wait time. Furthermore, the average wait time is tied to the number of CRPs of a matching topology, where a single CRP achieves minimum average wait time. In [3], the authors introduce necessary and sufficient constraints to ensure that the matching topology $\boldsymbol{M}$ induces a single CRP component. We adopt these constraints in order to achieve minimum wait time which we discuss next.

3.3.4 Optimization Formulation

Suppose the joint distribution of X, $Y(r)\ \forall r \in \mathcal{R}$ is known. Given a partitioning function p to assign individuals to queues, problem (3.1) can be solved via the MIO below.

$$\max \sum_{q\in\mathcal{Q}} \sum_{r\in\mathcal{R}} \tau_{qr} f_{qr} \tag{3.2a}$$

$$\text{s.t. } f_{qr}, \nu_{qr} \in \mathbb{R}_+, \gamma_r, \theta_q \in \mathbb{R} \qquad \forall q \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2b}$$

$$M_{qr}, z_{qr} \in \{0,1\} \qquad \forall q \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2c}$$

$$g_{qr}^{(k)} \in \mathbb{R}_+ \qquad \forall q, k \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2d}$$

$$\sum_{q\in\mathcal{Q}} f_{qr} = \mu_r \qquad \forall r \in \mathcal{R} \tag{3.2e}$$

$$\sum_{r\in\mathcal{R}} f_{qr} = \lambda_q \qquad \forall q \in \mathcal{Q} \tag{3.2f}$$

$$f_{qr} \le \lambda_q \mu_r \left(\theta_q + \gamma_r + \nu_{qr}\right) + Z(1 - m_{qr}) \qquad \forall q \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2g}$$

$$f_{qr} \ge \lambda_q \mu_r \left(\theta_q + \gamma_r + \nu_{qr}\right) - Z(1 - m_{qr}) \qquad \forall q \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2h}$$

$$f_{qr} \le Z m_{qr} \qquad \forall q \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2i}$$

$$f_{qr} \le Z z_{qr} \qquad \forall q \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2j}$$

$$\nu_{qr} \le (|\mathcal{Q}| + |\mathcal{R}| + 1) W (1 - z_{qr}) \qquad \forall q \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2k}$$

$$\sum_{q\in\mathcal{C}} g_{qr}^{(k)} = \mu_r \qquad \forall r \in \mathcal{R}, k \in \mathcal{Q} \tag{3.2l}$$

$$\sum_{r\in\mathcal{R}} h_{qr}^{(k)} = \lambda_q - \frac{\delta}{|\mathcal{Q}| - 1} \qquad \forall q \in \mathcal{Q} \setminus \{k\}, k \in \mathcal{Q} \tag{3.2m}$$

$$\sum_{r\in\mathcal{R}} g_{qr}^{(k)} \le Z m_{qr} \qquad \forall q, k \in \mathcal{Q}, r \in \mathcal{R} \tag{3.2n}$$

$$\sum_{r\in\mathcal{R}} g_{qr}^{(k)} = \lambda_k + \delta \qquad \forall q, k \in \mathcal{Q} \tag{3.2o}$$

$$\boldsymbol{F} \in \mathcal{F}. \tag{3.2p}$$

In this formulation, $\delta := \left(\prod_{q\in\mathcal{Q}} v_q \prod_{r\in\mathcal{R}} v_r\right)^{-1}$ and $\frac{w_q}{v_q} = \lambda_q, \frac{w_r}{v_r} = \mu_r$ are rational number representations. The constants W and Z are defined as: $W := 1/2 \max\{\max_{q\in\mathcal{Q}} 1/\lambda_q, \max_{r\in\mathcal{R}} 1/\mu_r\}$, and $Z := \max_{q\in\mathcal{Q}} \lambda_q \max_{r\in\mathcal{R}} \mu_r \left(\sum_{q\in\mathcal{Q}} 1/\lambda_q + \sum_{r\in\mathcal{R}} 1/\mu_r + (|\mathcal{Q}| + |\mathcal{R}| + 1)^2 W\right)$. Constraints (3.2e) and (3.2f) are the flow balance constraints. Constants W, Z ensure that constraints (3.2g)-(3.2k) impose the KKT conditions of the quadratic program that approximates steady-state-flow for a matching topology $\boldsymbol{M}$. Constraints (3.2l)-(3.2o) enforce a single CRP component to ensure minimum wait time. Finally, constraint (3.2p) is a fairness constraint where we can use any of the aforementioned examples. In order to solve problem (3.2), there are several parameters that must be estimated. In particular, we need to estimate τ_{qr} and $\boldsymbol{\lambda}$ which depend of p, as well as $\boldsymbol{\mu}$.

3.4 Solution Approach

We first partition $\mathcal{X}$ and then estimate CATE in each subset of the partition. We propose to use causal trees to achieve both tasks simultaneously [183]. Causal trees estimate CATE of binary interventions by partitioning the feature space into sub-populations that differ in the magnitude of their treatment effects. The method is based on regression trees, modified to estimate the goodness-of-fit of treatment effects. A key aspect of using causal trees for partitioning is that the cut points on features are such that the treatment effect variance within each leaf node is minimized. In other words, individuals who are similar in the treatment effect are grouped together in a leaf node. This results in queues that are tied to the treatment effect of resources which will result in improved policy value (see Section 3.5).

3.4.1 Assumptions

Causal trees rely on several key assumptions which are standard in causal inference for treatment effect estimation [88]. These assumptions are usually discussed for the case of binary treatments. Below, we provide a generalized form of the assumptions for multiple treatments.

Assumption 4 (Stable Unit Treatment Value Assumption (SUTVA))**.** *The treatment that one unit (individual) receives does not change the potential outcomes of other units.*

Assumption 5 (Consistency)**.** *The observed outcome agrees with the potential outcome under the treatment received.*

The implication of this assumption is that there are no different forms of each treatment which lead to different potential outcomes. In the housing allocation setting, this requires that there is only one version of PSH, RRH and SO.

Assumption 6 (Positivity)**.** *For all feature values, the probability of receiving any form of treatment is strictly positive, i.e.,*

$$\mathbb{P}(R = r|\boldsymbol{X} = \boldsymbol{x}) > 0 \; \forall r \in \mathcal{R} \; \boldsymbol{x} \in \mathcal{X}.$$

Th positivity assumption states that any individual should have a positive probability of receiving any treatment. Otherwise, there is no information about the distribution of outcome under some treatments and we will not be able to make inferences about it. In Section 3.5, we discuss the implications of this assumption in the context of HMIS data.

Assumption 7 (Conditional Exchangeability)**.** *Individuals receiving a treatment should be considered exchangeable, with respect to potential outcomes $Y(r), r \in \mathcal{R}$, with those not receiving it and vice versa. Mathematically,*

$$Y(1), \ldots, Y(|\mathcal{R}|) \perp R|\boldsymbol{X} = \boldsymbol{x} \; \forall \boldsymbol{x} \in \mathcal{X}.$$

Conditional exchangeability means that there are no unmeasured confounders that are a common cause of both treatment and outcomes. If unmeasured confounders exist, it is impossible to accurately estimate the causal effects. In experimental settings, conditional exchangeability is obtained through stratified randomization. In observational settings, however, a decision-maker only

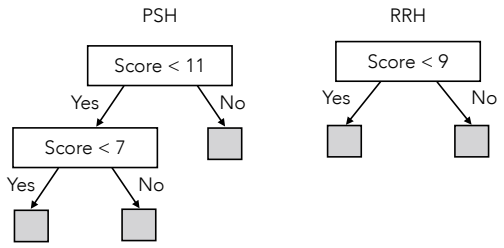


Figure 3.2: Example partitioning by sample causal trees for PSH and RRH interventions.

relies on passive observations. As a result, in order to increase the plausibility of this assumption, researchers typically include as many features as possible in $\boldsymbol{X}$ to ensure that as many confounders as possible between treatment and outcome are accounted for. In the housing allocation setting, the HMIS database contains a rich set of features (54 features) associated with different risk factors for homelessness which we use in order to obtain the treatment effect estimates. In Section 3.6, we discuss the consequences of violating the above assumptions.

3.4.2 Building the Partitioning Function

Next, we describe our approach for estimating CATE. We first consider a simple case with binary treatments, i.e., $|\mathcal{R}| = 2$ as causal trees work primarily for binary treatments. After training the causal tree using the data on a pair of treatments, the leaves induce a partition on the feature space $\mathcal{X}$. Hence, we can view the causal tree as the partitioning function p where each individual is uniquely mapped to a leaf node, i.e., a queue.

Extending to the case of $|\mathcal{R}| > 2$ is non-trivial. Assuming $r = 1$ is the baseline intervention, we construct $|\mathcal{R}| - 1$ separate causal trees to estimate CATE for $r \in \mathcal{R} \setminus \{1\}$. We denote the resulting causal trees or partitioning functions $p_r : \mathcal{X} \to \mathcal{Q}\ \forall r \in \mathcal{R} \setminus \{1\}$. We define $\mathcal{X}_r(q) = \{\boldsymbol{x} \in \mathcal{X} : p_r(\boldsymbol{x}) = q\}\, \forall r \in \mathcal{R} \setminus \{1\}, q \in \mathcal{Q}$ as the set of all individuals who belong to queue q according to partitioning function p_r. Also, let $q_r = p_r(\boldsymbol{x})$. In order to aggregate the individual partitioning functions to obtain a unified partition on $\mathcal{X}$, we consider the intersections of $\mathcal{X}_r(q)$

created by each tree. We define subsets $\mathcal{X}(q_1, \ldots, q_{|\mathcal{R}|-1}) = \bigcap_{r=1}^{|\mathcal{R}|-1} \mathcal{X}_r(q_r)$ for all combinations of $q_r \in \mathcal{Q}$. We can view $\mathcal{X}(q_1, \ldots, q_{|\mathcal{R}|-1})$ as a new (finer) partition on $\mathcal{X}$. We illustrate with an example using the housing allocation setting. Suppose we have constructed two causal trees for PSH and RRH according to Figure 3.2 such that PSH tree splits the vulnerability score into intervals of $[0, 6], [7, 10], [11, 17]$ and RRH creates $[0, 8], [9, 17]$ subsets. According to our procedure, the final queues are constructed using the intersection of these subsets. In other words, we obtain $[0, 6], [7, 8], [9, 10], [11, 17]$ which corresponds to four queues. We note that the granularity of the partition can be controlled through the tree depth or the minimum allowable number of data points in each leaf, both of which are adjustable parameters.

Finally, in order to estimate τ_{qr}, we should avoid using the estimates from each individual tree. The reason is that each tree estimates $\mathbb{E}[Y(r) - Y(1)|p(\boldsymbol{X}) = q, R \in \{1, r\}]\ \forall r \in \mathcal{R} \setminus \{1\}$. That is, a subset of the data associated with a pair of treatments is used to build each tree. Therefore, τ_{qr} values are not generalizable to the entire population and need to be re-evaluated over all data points that belong to a subset. We adopt Doubly Robust estimator (DR) for this task. Proposed in [62], DR combines an outcome regression with a model for the treatment assignment (propensity score) to estimate treatment effects. DR is an unbiased estimate of treatment effects, if at least one of the two models are correctly specified. Hence, it has a higher chance of reliable inference. CATE estimates $\hat{\tau}_{qr}$ are provided below.

$$\hat{\tau}_{qr} = \frac{1}{|\mathcal{I}_q|} \sum_{i \in \mathcal{I}_q} \left(\hat{y}(\boldsymbol{X}_i, r) + (Y_i - \hat{y}(\boldsymbol{X}_i, R_i)) \frac{\mathbb{I}(R_i = r)}{\bar{\pi}(R_i|\boldsymbol{X}_i)} \right) - \frac{1}{|\mathcal{I}_q|} \sum_{i \in \mathcal{I}_q} \left(\hat{y}(\boldsymbol{X}_i, 1) + (Y_i - \hat{y}(\boldsymbol{X}_i, R_i)) \frac{\mathbb{I}(R_i = 1)}{\bar{\pi}(R_i|\boldsymbol{X}_i)} \right),$$

where $\mathcal{I}_q := \{i \in \{1, \ldots, N\} : p(\boldsymbol{X}_i) = q\}$ is the set of indices in the historical data that belongs to q. Further, $\hat{y}$ and $\bar{\pi}$ are the outcome and historical policy (i.e., propensity score) models, respectively.

We end this section by discussing a practical consideration which is a desire to design policies that depend on low-dimensional features, such as risk scores. In cases that we only use risk scores,

not the full feature vector, it is critical that they satisfy the causal assumptions. We provide a risk score formulation that satisfies this requirement.

Proposition 9. *We define risk score functions as* $S_r(\boldsymbol{x}) = \mathbb{P}[Y(r) = 1|\boldsymbol{X} = \boldsymbol{x}]\ \forall r \in \mathcal{R}$. *Suppose* $\boldsymbol{S} \in \mathcal{S}$ *is a (random) vector of risk scores. Also, let* $\boldsymbol{Y} = (Y(1), \ldots, Y(|\mathcal{R}|))$ *be the vector of potential outcomes. The following statements hold:*

1. *If* $\boldsymbol{Y} \perp R|\boldsymbol{X}$, *then* $\boldsymbol{Y} \perp R|\boldsymbol{S}$.

2. *If* $\mathbb{P}\left(\mathbb{P}(R = r|\boldsymbol{X} = \boldsymbol{x}) > 0\right) = 1\ \forall \boldsymbol{x} \in \mathcal{X}$, *then* $\mathbb{P}\left(\mathbb{P}(R = r|\boldsymbol{S} = \boldsymbol{s}) > 0\right) = 1\ \forall \boldsymbol{s} \in \mathcal{S}$.

Under causal assumptions, $S_r(\boldsymbol{x}) = \mathbb{P}(Y(r) = 1|\boldsymbol{X} = \boldsymbol{x}, R = r) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}, R = r)$, which relies on observed data, rather than counterfactuals. According to Proposition 9, as in general individuals respond differently to various treatments, one risk score per resource may be required in order to summarize the information of $\boldsymbol{X}$. Alternatively, one can utilize all features to learn the propensity and outcome models and use those estimates in causal tree construction.

3.5 Computational Results

We conduct two sets of experiments to study the performance of our approach to design resource allocation policies: *(i)* synthetic experiments where the treatment and potential outcomes are generated according to a known model; *(ii)* experiments on the housing allocation system based on HMIS data for youth experiencing homelessness. We use the causal tree implementation in the `grf` package in R. We control the partition granularity by changing the *minimum node size* parameter which is minimum number of observations in each tree leaf. We evaluate policies using three estimators from the causal inference literature [62]: Inverse Propensity Weighting (IPW) which corrects the mismatch between the historical policy and new policy by re-weighting the data points with their propensity values, Direct Method (DM) which uses regression models to estimate the unobserved outcomes, and DR. In addition, we include objective value of Problem (3.2)

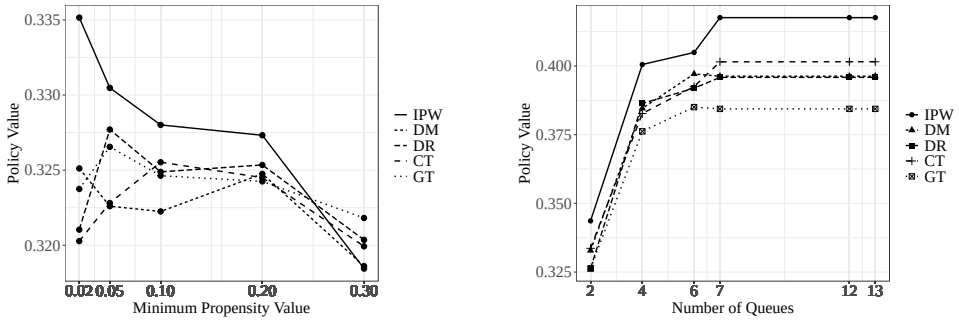


Figure 3.3: Synthetic data experiments: policy value vs. the minimum propensity weight (left) and policy value vs. the number of queues (right). Each line corresponds to a different estimator.

obtained by matching flow and CATE estimates (CT). When models of outcome and propensity are correctly specified, the above estimators are all unbiased [62].

3.5.1 Synthetic Experiments

We generate synthetic potential outcomes and resource assignments in the HMIS data collected between 2015 and 2017 from 16 communities across the United States [46]. We use the following setting using vulnerability score S (unless mentioned otherwise): $\bar{\pi}(\text{SO}|S > 0.2) = 0.3$, $\bar{\pi}(\text{SO}|0.0 < S \leq 0.2) = 0.3$ and $\bar{\pi}(\text{SO}|S \leq 0.0) = 0.3$. Additionally, $\bar{\pi}(\text{RRH}|S > 0.2) = 0.2$, $\bar{\pi}(\text{RRH}|0.0 < S \leq 0.2) = 0.4$ and $\bar{\pi}(\text{RRH}|S \leq 0.0) = 0.3$ and finally, $\bar{\pi}(\text{PSH}|S > 0.2) = 0.5$, $\bar{\pi}(\text{PSH}|0.0 < S \leq 0.2) = 0.3$ and $\bar{\pi}(\text{PSH}|S \leq 0.0) = 0.4$. The potential outcomes are sampled from binomial distributions with probabilities that depend on S. For PSH, we use $\mathbb{E}[Y(\text{PSH})|S \leq 0.3] = 0.6$, $\mathbb{E}[Y(\text{PSH})|0.3 < S \leq 0.5] = 0.2$ and $\mathbb{E}[Y(\text{PSH})|0.5 < S] = 0.6$. For RRH, $\mathbb{E}[Y(\text{RRH})|S \leq 0.2] = 0.2$, $\mathbb{E}[Y(\text{RRH})|0.2 < S \leq 0.7] = 0.6$ and $\mathbb{E}[Y(\text{RRH})|0.7 < S] = 0.2$. Finally, $\mathbb{E}[Y(\text{SO})] = 0$. We evaluate policies obtained by solving Problem (3.2). We use decision trees for outcome and propensity score models.

One of the goals of the synthetic experiments is to compare different estimators in a setting where we observe the potential outcomes. Specifically, we study the performance of the estimators for policy evaluation when propensity values are varied. We generate different datasets by changing

the propensity values $\bar{\pi}(\text{PSH}|0.0 < S \leq 0.2) = \alpha$ and $\bar{\pi}(\text{RRH}|0.0 < S \leq 0.2) = 0.7 - \alpha$ for $\alpha \in \{0.02, 0.05, 0.1, 0.2, 0.3\}$ and obtain the optimal policy for each dataset. Figure 3.3 shows optimal policy values according to different estimators. We observe that across the $x-$axis range, DR, DM and CT result in similar estimates which also agrees with the ground truth (GT). However, when the minimum propensity score is small (< 0.05), IPW diverges from GT. This is consistent with other findings in the literature suggesting that when propensities are too close to 0 or 1, non-parametric estimators tend to have higher variance and converge at a slower rate (with the number of data points) [108].

Next, we investigate the effect of treatment heterogeneity on the value of the optimal policy. In particular, we study how much the granularity of partitions, or the number of queues, impacts the policy value. Figure 3.3 summarizes the results. When the number of queues is equal to 1, the optimal policy is at its minimum value. In this case, the policy corresponds to an FCFS policy as individuals queue in a single line and are prioritized according to the their arrival times. The optimal policy value gradually increases ($\sim 25\%$ according to GT) as the number of queues increases until it flattens. This suggests that by increasing the number of queues, we can leverage the treatment effect heterogeneity across the queues to allocate resources more efficiently.

3.5.2 HMIS Data of Youth Experiencing Homelessness

We now showcase the performance of our approach to design policies that allocate resource among the U.S. homeless youth. We defer the details on data preparation to the Appendix.

3.5.3 Data Pre-Processing and Estimation

Outcome Definition. We focus on the likelihood of stable exit from homelessness.

An exit from the system can be to any of the following destinations: "family," "self-resolved," "RRH," "PSH," "deceased," or "incarcerated." Exiting due to incarceration or being deceased are undesirable outcomes and are encoded as $Y = 0$ (left branch). "Family," "self-resolve," "RRH,"

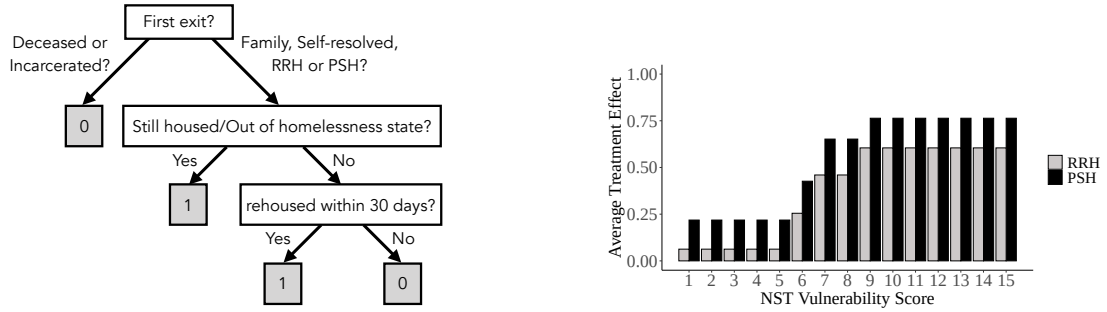


Figure 3.4: HMIS data: success definition flow chart (left) and heterogeneous treatment effect using DR method (right)

and "PSH" are desirable outcomes but may be temporary exits, meaning that the individual may return to homelessness shortly after. In addition, there are recorded exits that are simply due to a "move" in the system from one service to another. We distinguish between these cases by checking whether an individual is "still housed", i.e., is at the stable exit destination. If re-housed, we consider a 30-day threshold to decide whether it is a return to homelessness ($Y = 0$) or a move in the system ($Y = 1$). This procedure for defining outcome is summarized in Figure 3.4.

Propensity Estimation. In order to obtain an unbiased estimation of the policy value, IPW and DR approaches rely on propensity values. In our setting, the propensities are unknown but can be estimated from data. This poses a challenge to find a model that fits the data while being well-calibrated. We use different statistical models for multi-class classification to estimate $\mathbb{P}(R = r|\boldsymbol{X} = \boldsymbol{x}), \forall r \in \mathcal{R}$. We evaluate models based on the predictive power, calibration, and fairness. For fairness, we adopt the test fairness criteria in [51] since evaluating the policy value across different protected groups requires propensity values that are well-calibrated for those groups. We defer the details on model selection to the Appendix. We note that the original dataset does not satisfy the positivity assumption. That is, some groups of individuals have only received a subset of the resources. Therefore, for data points with propensities less than 0.001, we follow the status-quo policy and we exclude them from the policy optimization.

Outcome Estimation. DM and DR methods rely on a model of the outcome under different resources. We compare an array of models in terms of accuracy, calibration, and test-fairness. The results are summarized in the Appendix.

Heterogenous Treatment Effect Estimation. We use causal trees with minimum node size equal to 15 to estimate the average treatment effects across the NST score range for RRH and PSH. According to Figure 3.4, PSH consistently has a higher treatment effect than RRH indicating that it is a more effective resource. Further, the treatment effect of both resources increase with score which suggests that higher-scored individuals benefit more from these resources. We also provide results on the (unbiased) probability of exiting homelessness versus NST score in the Appendix.

Arrival Rate. Once the queues are constructed, we can estimate the arrival rate of individuals from data. Given the heavy-traffic condition, we calculate the required rate of SO as $\mu_{\text{SO}} = \max(\lambda_{\mathcal{Q}} - \mu_{\text{RRH}} - \mu_{\text{PSH}}, 0)$.

3.5.4 Policy Optimization Results

We now present the policy optimization results along three distinct objectives: policy value measured in terms of rate of stable exit from homelessness, fairness by race and age, and wait time. Table 3.1 summarizes the results, where OPT is the optimal policy value without fairness constraints and OPT-fair (race), and OPT-fair (age) represent our method with fairness constraints over race and age, respectively. As baselines, we simulate both a fully FCFS policy and the status quo policy SQ (see Figure 3.1). We also compare with the deployed policy in the data SQ (data). As IPW suffers in small-propensity settings, we exclude it from the estimators.

From Table 3.1, OPT, OPT-fair (race), and OPT-far (age) all outperform the baseline policies. Specifically, OPT significantly improves the rate of stable exit from homelessness by 19% and 13% (under DR estimates) over SQ and FCFS policies, respectively. Perhaps surprisingly, SQ performs worse than FCFS which is due to how the cut scores are designed. Specifically, according to SQ individuals with scores 4-7 are matched to RRH. However, the RRH treatment effect is highest

Policy	Rates of Stable Exit from Homelessness			Wait Time (days)
	CT	DM	DR	
OPT	0.76	0.74	0.75	142.67
OPT-Fair (race)	0.76	0.75	0.76	142.64
OPT-Fair (age)	0.76	0.75	0.75	142.64
FCFS	0.68	0.68	0.66	142.64
SQ	0.66	0.63	0.63	182.21
SQ (data)	0.73	0.73	0.73	156.77

Table 3.1: Out-of-sample estimated policy performance measured in terms of rates of stable exit from homelessness and wait times.

for scores above 7 (See Figure 3.4). Compared to SQ (data), our policy values are competitive. We improve the wait time over SQ and SQ (data) by 21% and 9%, respectively and obtain values similar to FCFS policy, suggesting that further algorithmic improvement is not possible unless problem inputs, such as resource arrival rates, change.

Figure 3.5 compares the worst-case rate of exiting homelessness across age (below and over 17 years old) and racial groups (White, Black, and Other) according to DR estimator. First, we observe that an FCFS policy does not necessarily result in policies that are fair in terms of their outcomes neither by age nor by race. This is because FCFS policies ignore treatment effect heterogeneity. In other words, according to the FCFS discipline, everyone has the same probability of receiving any one of the resource types (fairness in allocation). However, not everyone benefits equally from the resources. Indeed, Black individuals seem to suffer the most under a fully FCFS policy. SQ also yields a low worst-case performance mainly due its low overall performance. SQ (data) has relatively better worst-case performance. However, there is

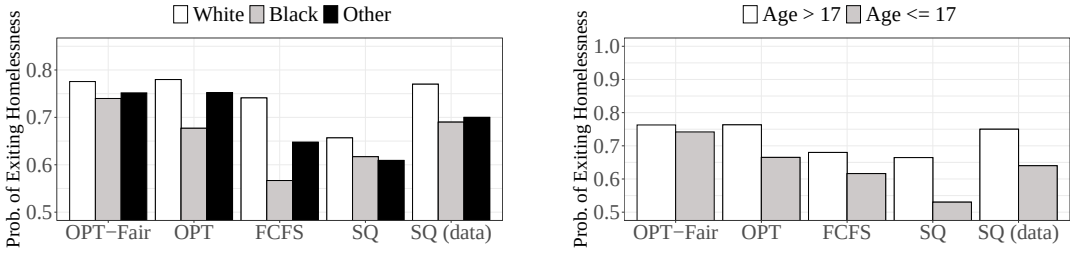


Figure 3.5: Out-of-sample rates of exit from homelessness by race (left panel) and age (right panel) using the DR estimator.

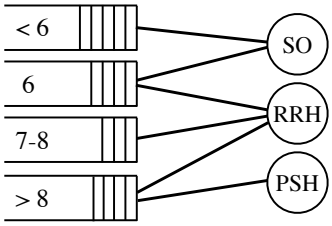


Figure 3.6: Optimal Topology

still a significant gap between the performance of Black/Other groups and Whites. By explicitly imposing fairness constraints on policy outcomes across protected groups, OPT-Fair significantly improves the performance for the Black and Other groups. Figure 3.5, similar observations can be made for fairness by age, where compared to baselines with no fairness considerations, OPT-Fair exhibits significant improvements in the policy value for those with age below 17.

We now present a schematic diagram of OPT and OPT-fair matching topologies. Figure 3.6 is the matching topology corresponding to OPT policy. Compared to SQ, OPT uses different cut points on NST score, specifically for the lower-scoring individuals. Across the four score groups, we observe a gradual transition from eligibility for a more resource-intensive intervention (PSH) to a basic intervention (SO). Figure 3.7 depicts OPT-fair topology for fairness on race, in which

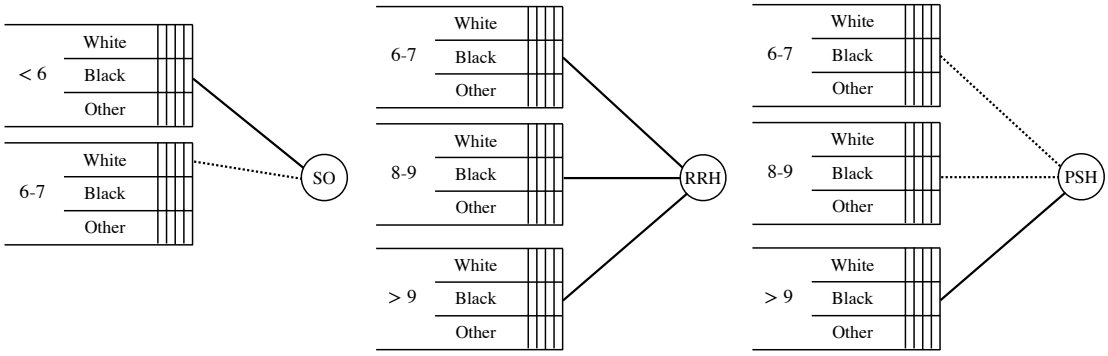


Figure 3.7: Matching topology split by resource type: left (SO), middle (RRH) and right (PSH). Individuals are divided into four different score groups: $S < 6$, $S = \{6, 7\}$, $S = \{8, 9\}$, $S > 9$. Queues are constructed based on score groups and race jointly. Solid lines indicate that a resource is connected to the entire score group (a collection of queues). Dotted lines indicate connection to a single queue within the score group. For example, in the left figure, SO is only connected to the individuals with $S = \{6, 7\}$ and race White.

queues are constructed using the joint values of NST score and race. According to this figure, PSH is matched to all individuals with scores above 9 as well as mid-scoring Black individuals, i.e., $6 \leq$ score ≤ 9. RRH is connected to every individual in the mid-score range. Our modeling strategy uses the protected characteristics in order to ensure fairness. This is motivated by discussions with our community advisory board, including housing providers/matchers and people with past history of homelessness, who suggested that in order to create a fair housing allocation system there ought to be special accommodations for historically disadvantaged people. Our policies align with affirmative action policies that take individuals' protected attributes into account in order to overcome present disparities of past practices, policies, or barriers by prioritizing resources for underserved or vulnerable groups. In this regard, recently HUD restored Affirmatively Furthering Fair Housing rule that requires "HUD to administer its programs and activities relating to housing and urban development in a manner that affirmatively furthers the purposes of the Fair Housing Act", extending the existing non-discrimination mandates [58].

Our approach can also be extended to non-affirmative policies. This is possible by imposing constraints that ensure a topology has the same connections to all protected groups within a score group. Such constraints are expressible as linear constraints and can be easily incorporated in Problem (3.2). We demonstrate the result for fairness on race in Figure 3.8. We observe that all individuals who belong to a certain queue, regardless of their race, are eligible for the same types of resources. However, as a result of combining the queues, the worst-case policy value across the racial groups decreases from 0.76 to 0.73 which still outperforms SQ and SQ (data)

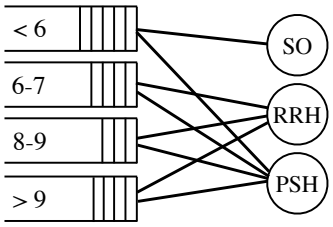


Figure 3.8: Fair topology (race)

with worse-case value of 0.61 and 0.69, respectively. We defer the results for fairness by age to Appendix C.

3.6 Conclusion and Broader Impact

Recently, there has been a significant growth in algorithms that assist decision-making across various domains [159, 82, 168, 131]. Homelessness is a pressing societal problem with complex fairness considerations which can benefit greatly from data-driven solutions. As empirical evidence on ethical side effects of algorithmic decision-making is growing, care needs to be taken to minimize the possibility of indirect or unintentional harms of such systems. We take steps towards this goal. Specifically, we propose *interpretable* data-driven policies that make it easy for a decision-maker to identify and prevent potential harms. Further, we center our development around issues of *fairness* that can creep into data from different sources such as past discriminatory practices. We provide a flexible framework to design policies that overcome such disparities while ensuring efficient allocations in terms of wait time and policy outcome.

There are also crucial consideration before applying our framework in real-world. Our approach relies on several key assumptions about the data. Specifically, the consistency assumption requires that there is only one version of PSH, RRH, and SO. In practice, different organizations may implement different variants of these interventions. For example, combining substance abuse intervention with PSH and RRH. Such granular information about the interventions, however, is not currently recorded in the data which may impact CATE estimates. Further, the exchangeability assumption requires that there are no unobserved confounders between treatment assignment and outcomes. Even though our dataset consists of a rich set of features for each individual, in practice, unobserved factors may influence the allocation of resources which calls for more rigorous inspection of service assignment processes. Unobserved confounders may lead to biased estimates of treatment effects which in turn impacts the allocation policies. In addition, our dataset consists

of samples from 16 communities across the U.S., which may not be representative of new communities or populations. Hence, the external validity of such policies should be carefully studied before applying to new populations. Finally, there are other domain-specific constraints that we have not considered as they require collecting additional data. For example, resources can not be moved between different CoCs. We leave such considerations to future work.

Chapter 4

Causal Inference for Ethical Decision-Making

4.1 Introduction

Recently, there has been a growing interest in applying causality for unfairness evaluation and mitigation [115, 109, 134, 50]. Causality provides a conceptual and technical framework for addressing questions about the effect of (hypothetical) interventions on, in this context, sensitive attributes such as race, gender, etc. This is in contrast with fairness criteria that merely rely on passive observations [44, 101, 85, 191, 110]. Observational criteria achieve fairness by constraining the relationships between variables, often in conflicting ways. Consequently, it has been shown that it is impossible to satisfy these criteria simultaneously on a dataset [110, 51, 169]. Causality helps unify these different perspectives by shifting the focus from association to causation in order to identify and mitigate the *sources of disparity*. This perspective is also more compatible with legal requirements of evaluating algorithmic bias discussed in earlier work [186].

Nevertheless, causal fairness too has been subject to criticism. One objection is around the validity of the assumptions in causal modeling. The majority of recent research on causal fairness has focused on structural causal models, which encode the relationships between variables via a Directed Acyclic Graph (DAG) [115, 134, 50]. In realistic settings, however, constructing the DAG model is a challenging task. In particular, it is generally difficult to come up with arguments for the absence of links without conducting controlled experiments [95]. Causal discovery from observational data also relies on strong untestable assumptions or do not generally pin down all possible causal details in a DAG [170, 120].

There are also concerns about considering categories such as race or gender as a cause [122, 112, 92, 104]. From one perspective, most of these attributes are determined at the time of an individual's conception and are modeled as source nodes in a causal graph which can directly or indirectly influence the descendent variables. This view raises several major issues. Through such conceptualization, in order to evaluate and mitigate unfairness, one is inevitably required to

identify all possible pathways through which sensitive attributes influence an outcome. In addition to the modelling challenge this view poses, in practice a single entity may not be held liable for the discrimination across an entire causal pathway. In this regard, many anti-discrimination mechanisms investigate whether a *particular person or institutional actor* has behaved in a discriminatory manner. For example, in the employment setting, a racial discrimination lawsuit aims to determine whether a firm has withheld some benefits such as hiring with regard to racial identity of the applicant. However, disparities in hiring rates for different groups might be a reflection of either discrimination or differences in the applicant pool's qualifications. For example, if past discrimination in the educational system has led to some applicants having lower educational achievements, by hiring based on educational achievements, the employer will perpetuate the effects of this discrimination. Under anti-discrimination law, however, as long as the employer makes the hiring decision based on educational achievements that are legitimately connected with the job and business needs—with no regard to race—no liability is attached. In fact, if the employer seeks to proactively address past societal discrimination, this could lead to reverse discrimination lawsuits [129]. Another issue is post-treatment bias, which arises when one controls for post-treatment variables, resulting in biased estimates of the treatment effect [160]. Since some attributes such as race, gender, etc. are fixed at the time of one's conception, almost all measurable variables become post-treatment. Hence, conditioning on those variables may lead to misleading estimates of discrimination. Removing those variables, e.g., as proposed in [115, 134], leaves little to no information for valid causal analysis.

Alternatively, many view attributes such as race or gender as social constructs that evolve over the course an individual's life. Recently, [93, 104] studied epistemological and ontological aspects of counterfactuals in the context of fairness evaluation. In [93], the authors argue that social categories such as race may not admit counterfactual manipulation. In [104], the authors aim to address this problem by proposing a set of tenets which require a decision-maker to state implicit and unspecified assumptions about social ontology as explicitly as possible. Despite recent efforts,

there has been limited empirical investigation on how the nature of the intervention impacts the scope and validity of causal analysis of sensitive attributes and conclusions one draws.

In this work, we investigate the practical and epistemological challenges of applying causality for fairness evaluation. In particular, we highlight two key aspects that are often ignored in the current causal fairness literature: *nature* and *timing* of the interventions on social categories such as race, gender, etc. Further, we discuss the impact of this specification on the plausibility of causal assumptions. To facilitate this discussion, we draw a distinction between intervening on immutable attributes and their perception, and demonstrate how such conceptualization allows us to disentangle the potential unfairness along causal pathways and attribute it to the respective actors. The idea that perceptions matter and can be manipulated is not new. For example, researchers have examined the effect of manipulated names associated with political speeches [164] and resumes [29]. Nevertheless, in the machine learning literature, little attention has been paid to the consequences for valid causal inference for unfairness evaluation and mitigation. We make the following contributions:

- We propose a causal framework to investigate and mitigate unfairness of a particular actor's behavior, along a causal pathway. To the best of our knowledge, no prior work has aimed to isolate such effects for fair prediction. To tackle this problem, we highlight the importance of identifying the timing and nature of the intervention on social categories and its impact on conducting valid causal analysis including avoiding post-treatment bias.

- We illustrate how causality can address the limitations of existing fairness criteria, including those that depend upon statistical correlations. In particular, we introduce the causal variants of the popular statistical criteria for fairness and we make a novel observation that under the causal framework there is indeed no fundamental disagreement between different fairness definitions.

- We conduct extensive experiments where we demonstrate the effectiveness of our methodology for unfairness evaluation and mitigation compared to common baselines. Our results indicate that the causal framework is able to effectively identify and remove disparities at various stages of decision-making.

4.2 Related Work

There are two main frameworks for causal inference: structural causal models [90], also referred to as DAGs, and the potential outcomes framework (POF) [162]. DAGs can be viewed as a sequence of steps for generating a distribution from independent noise variables. Causal queries are performed by changing the value of a variable and propagating its effect through the DAG [90]. POF, on the other hand, starts by defining the counterfactuals with reference to an *intervention* and postulates potential outcomes under different interventions, albeit some unobserved. In general, DAGs encode more assumptions about the relationships of the variables; i.e., one can derive potential outcomes from a DAG, but potential outcomes alone are not sufficient to construct the DAG. Consequently, POF has been more widely adopted in empirical research, including bias evaluation outside of ML [29, 179]. More detailed discussion on the differences between the two frameworks in relation to empirical research can be found in [95]. Causal inference on immutable attributes has appeared in several works including [179, 109] via proxy variables and [83] through the perception of an immutable attribute. In this work, we follow the footsteps of [83] and provide a rigorous framework to reason about the causal effect of immutable attributes which helps avoid some of the common issues in causal inference including post-treatment bias.

Recently, there has been much interest in causality in the machine learning community, where the majority of works have adopted the DAG framework [115, 109, 194, 193, 119, 50] with a few exceptions that rely on POF [134, 107]. Specifically, [115] provides an individual-based causal fairness definition that renders a decision fair towards an individual if it is the same in the actual

world and a counterfactual world where the individual possessed a different sensitive attribute. In [109], the authors propose *proxy discrimination* as (indirect) discrimination via proxy variables such as name, visual features, and language which are more amenable to manipulation. Additionally, [134, 50] study path-specific discrimination, where the former proposes to remove the descendants of the protected attribute under the unfair pathway and the latter aims to correct the those variables. In [107], the authors propose two causal definitions of group fairness: fair on average causal effect (FACE), and fair on average causal effect on the treated (FACT) and show how these quantities can be estimated for specific attributes such as race or gender as the treatment. The authors restrict their attention to the fairness evaluation task and do not discuss the distinction between pre- and post-treatment variables. Further, [194, 193] discusses counterfactual direct, indirect, and spurious effects and provides formulas to identify these quantities from observational data. These works rely on a causal model, or DAG, and develop different methodologies to identify and mitigate unfairness. However, a clear discussion of the causal assumptions is typically missing, which consequently hinders the adoption of these methods in practice. In addition, the validity of the causal assumptions are influenced by the nature of the postulated intervention and its timing, which is not clearly articulated in the current literature. In many applications, discrimination by specific individuals or institutional actors is the subject of a study not an entire causal pathway. Our work makes this distinction and discusses the importance of specifying the timing and nature of a hypothetical intervention to conduct such analyses.

Finally, we briefly review the observational notions of fairness. Demographic parity and its variants have been studied in numerous papers [63, 69, 54]. Also referred to as statistical parity, this fairness criteria requires the average outcome to be the same across different sensitive groups. Conditional statistical parity [69, 54] imposes a similar requirement after conditioning on a set of legitimate factors. In the classification setting, equalized odds and a relaxed variant, equality of opportunity, have been proposed [85] to measure the disparities in the error rate across different

sensitive groups. The aforementioned criteria can be expressed using probability statements involving the observed random variables at hand, hence the name observational. These criteria are often easy to state and interpret. However, they suffer from a major limitation: it is impossible to simultaneously achieve these criteria on any particular dataset [110, 51, 169]. In this work, we revisit these notions and introduce their causal variants, where we show that under the causal framework, there is no fundamental disagreement between different criteria.

4.3 Causal Fairness: A Potential Outcomes Perspective

We consider a decision-making scenario where $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^n$ is the available set of attributes for an individual which we aim to use in order to make a (discrete) decision $Y \in \{0, 1\}$. An individual is further characterized by a sensitive attribute $A \in \{0, 1\}$ for which fair treatment is important. We assume A is a single binary variable, however, our discussion can naturally be extended to cases where A has more than two levels. It also applies when there is more than one sensitive attribute, such as the intersection of race and gender, by considering their joint values. Causal fairness views the unfairness evaluation and mitigation problem as a counterfactual inference problem. For example, we aim to answer questions of type: *What would have been the hiring decision, if the person had been perceived to be of a different gender?* or *Would the person have been arrested if they had been perceived to be a different race?* Such causal criteria are centered around the notion of an *intervention* or *treatment* on social categories such as gender and race. Formally, we build on POF [162] and define $Y(A), A \in \{0, 1\}$ as random variables describing the potential outcomes under different values of A, i.e., the outcome after we manipulate one's sensitive attribute A (its perception). It is important to note that for any individual, only one of the values of $Y(A)$ is observed which is the outcome corresponding to the possessed value of A. Other outcomes are considered as counterfactual quantities and are treated as unobservable variables.

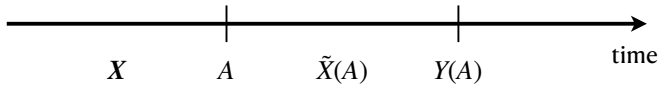


Figure 4.1: Decision-making timeline: the time when one's sensitive attribute A is perceived determines pre- and post-treatment variables. Here, $\boldsymbol{X}$ is the vector of pre-treatment variables, $\tilde{X}$ is a post-treatment variable and Y is the outcome or decision.

In this work, we take a decision-maker's perspective considering how their perception of one's sensitive attribute may lead to different decisions. Through this conceptualization, it is possible for discrimination to operate not just at one point in time and within one particular domain but at various points within and across multiple domains throughout the course of an individual's life. For example, in the context of racial discrimination, earlier work has recognized potential points of discrimination across different domains including labor market, education, etc. [147].

Consequently, we need to specify the point in time at which we wish to measure and mitigate unfairness. In causal terms, this is closely related to the notion of timing of the intervention, i.e., the time at which one's sensitive attribute is perceived by an actor. To illustrate, consider a hiring scenario and suppose we are interested in evaluating whether the hiring decision is fair with respect to gender or not. We can investigate unfairness at different stages, e.g., from the first time an individual comes into contact with the company (e.g., resume review), progresses in the system (during interviews), or when the final decision is being made. We may even take a much broader perspective and investigate the effect of gender from the point an individual attends college and study how gender affects education and subsequently the opportunities in the job market. Indeed, as we expand our view the causal inference problem we are faced with becomes increasingly more challenging but the conceptual framework remains valid.

Both timing and nature of the intervention impact the conclusions we draw. For example, under an unfair educational system, a hiring decision that is based on educational achievements will perpetuate those biases, even if it treats individuals fairly given their educational background. Similarly, a discriminatory interview process will result in an unfair hiring decision. However,

the difference is that in the latter, the company is now liable for the discriminatory behavior as it stems from the a point in its decision-making process. The timing of the intervention is thus important in conducting causal analysis. In particular, consider an interview process which is discriminatory, resulting in unfair interview scores for a particular group. In our fairness evaluation, if we control for the interview score, we will find no relationship between gender and hiring decision contrary to our intuition that the decision-maker discriminates between female and male candidates through the interview score. This observation is due to post-treatment bias cautioned in the causal inference literature which happens when variables that are fixed after the intervention are used in evaluating the treatment effect [132]. Figure 4.1 demonstrates this over a decision-making timeline. After we fix the point of (hypothetical) intervention on A, variables $\tilde{X} \in \tilde{\mathcal{X}} \subseteq \mathbb{R}^m$ determined afterwards are considered as post-treatment variables and in principle are affected by A. Hence, we introduce the counterfactual values of $\tilde{X}(0), \tilde{X}(1)$ to differentiate between pre-treatment and post-treatment variables. Consequently, the observed values of post-treatment variables are determined as $\tilde{X} = \tilde{X}(0)(1 - A) + \tilde{X}(1)A$.

Furthermore, the nature of the intervention influences the causal effect that we are able to uncover. For instance, in the study conducted in [29], the authors manipulated the names on the resumes to measure racial discrimination which only allowed them to capture the level of discrimination exhibited through the relationship between one's name and perception of race. Under a different manipulation, e.g., zip code of the applicant, the outcome of the study would have been different. In observational studies, where the analyst has no control over how an individual's sensitive attribute is perceived, a careful examination of mechanisms through which one's attributes are perceived is necessary. Indeed, it is possible to identify several mechanisms affecting perceived attributes (e.g., name, clothing, language, etc.). In this case, it is possible to study the joint effect of the mechanisms by modeling the missing counterfactual values, under each mechanism, as random variables with a distribution. The distribution for each individual's

missing counterfactual value can then be represented by a stochastic mixture of distributions associated with each mechanism [83].

Building on the above discussion, we define fairness in terms of the treatment effect of a *specific intervention* on perceived sensitive attribute at a *particular point in time*. We refer to this notion as causal parity and under the POF, we can express it mathematically via the following definition.

Definition 3 (Causal Parity)**.** *A decision-making process achieves causal parity if* $\mathbb{E}[Y(1) - Y(0)] = 0$.

In the above definition, $\tau = \mathbb{E}[Y(1) - Y(0)]$ is the treatment effect of A on Y. As stated earlier, both potential outcomes $Y(0), Y(1)$ are not simultaneously observed for any individual. In order to conduct meaningful causal inference to identify the treatment effects several assumptions are necessary. We review the assumptions and discuss how the precise specification of the intervention helps establish their plausibility.

4.3.1 Causal Assumptions for Identification

Assumption 8. *There is a set of established conditions under which causal inference becomes possible:*

- *Stable Unit Treatment Value Assumption (SUTVA): It states the treatment that one unit (individual) receives does not change the potential outcomes of other units.*

- *Consistency: Formally,* $Y = Y(0)(1 - A) + Y(1)A$. *In words,* Y *agrees with the potential outcome under the respective treatment. The implication of this assumption is that there are no two "flavors" or versions of treatment such that* $A = 1$ *under both versions but the potential outcome for* Y *would be different under the alternative versions.*

- *Positivity: At each level of pre-treatment variables $\boldsymbol{X}$, the probability of receiving any form of treatment is strictly positive. Mathematically,*

$$\mathbb{P}\left(\mathbb{P}(A = a \mid \boldsymbol{X} = \boldsymbol{x}) > 0\right) = 1 \; \forall a \in \{0, 1\}, \boldsymbol{x} \in \mathcal{X}.$$

- *Conditional Exchangeability: it states that those individuals receiving the treatment should be considered exchangeable (with respect to potential outcomes $\boldsymbol{Y}$ and the post-treatment variable $\tilde{X}$) with those not receiving the treatment and vice versa. Mathematically,*

$$\boldsymbol{Y}, \tilde{X} \perp A \mid \boldsymbol{X} = \boldsymbol{x} \; \forall \boldsymbol{x} \in \mathcal{X},$$

where $\boldsymbol{Y} = \{Y(0), Y(1)\}, \tilde{X} = \left\{\tilde{X}(0), \tilde{X}(1)\right\}$ and $\boldsymbol{X}$ is the vector of pre-treatment variables.

Earlier works have emphasized the criticality of these assumptions in determining the causal effects [163]. Here, we highlight their importance in the context of fairness evaluation. SUTVA can also be viewed as a non-interference assumption and depends very much on the problem under study and the choice of the decision-maker. For example, for a recruiter as the decider, one should think carefully whether the recruiter's decision to proceed with an application is independent from case to case. If a recruiter screened three candidates in a row with exceptional resumes, they might raise their standards when judging the fourth resume. In this case, SUTVA is violated as historical data on other candidates influences the future candidates outcomes.

The consistency assumption means, for example, that for candidates perceived as either male or female, an employer would not base the hiring decision on the level of "manliness." Similarly, the degree of "blackness" of an individual should not affect the decision made for an individual. This assumption, however, can be potentially relaxed with information beyond what is typically assumed. For example, if an accurate estimate of the level of "manliness" or skin color were recorded, then the treatment could be conceptualized as having multiple levels [83]. Consistency

can also be viewed as treatment invariance, which we discussed in the previous section in the context of nature of intervention on social categories. When intervening on social categories such as race, it is possible that different factors contribute to the perception of one's sensitive attribute. Under consistency, one needs to make sure that there is sufficient data in order to capture the different levels of "race." Without such nuanced data, it is still possible to measure the causal effect, but the interpretation changes, as the estimated causal effect is an average of multiple potential treatments.

The positivity assumption is also essential in order identify the treatment effect. It requires that there is not a complete overlap between the treatment assignment and pre-treatment variables. For example, if all of the women in a hiring pool have a PhD, and all of the men only have a Master's degree, then it is not possible to separate the effect of gender discrimination from the effect of the educational attainment on the employment decision. Positivity is often easy to verify from the data once the pre-treatment variables $\boldsymbol{X}$ are determined.

Conditional exchangeability is one of the cornerstone assumptions for causal inference, which is in principle impossible to verify in observational studies. Conditional exchangeability in experimental settings can be obtained through stratified randomization. In order to increase the plausibility of this assumption in observational contexts, analysts typically include as many pre-treatment variables as possible to ensure that as many confounders as possible between treatment and outcome are accounted for. Intuitively, the goal is to ensure that once all of the pre-treatment variables $\boldsymbol{X}$ are controlled for, the allocation of individuals between treatment and control is as close to random as possible. In the fairness setting, this would mean that, after controlling for $\boldsymbol{X}$, the only systematic difference between the two groups is the perception of their protected attribute (i.e., whether they were discriminated against), allowing for an empirical estimate of the effect of discrimination. We note that in the exchangeability assumption, we have the conditional independence of the counterfactuals of both $\tilde{X}$ and $\boldsymbol{Y}$. This is a key distinction with earlier work [107] that does not differentiate between pre- and post-treatment variables.

In more complicated settings, where an individual interacts with multiple parts of a system, we may have more than one choice of decision-maker to study. In such situations, an analyst may have to balance the need to make the exchangeability assumption plausible against the desire to study a decision-maker's behavior early in the decision-making chain. Choosing the timing of the intervention towards the later interactions renders more measured variables pre-treatment which in turn can make the exchangeability assumption more plausible. However, by treating such variables as pre-treatment and thus conditioning on them in the analysis, the analyst forgoes the detection of any prior discrimination that may have affected the values of these variables. In cases where there is sufficient data to detect discrimination starting from earlier stages of decision-making, it may be still important to pin down the different sources of discrimination throughout the decision-making process. For example, in the hiring context, suppose from the onset (the first interaction of the applicant with the company), a rich set of data about the applicant's background and qualifications is collected that allows an analyst to determine the hiring process is unfair towards to a group. In such a case, it is important to understand whether discrimination is attributed to the recruitment process, the interview stage or the final hiring process. Additionally, there may be a long delay between the time of perceiving an individual's sensitive attribute and outcome. In this case, it may be helpful to use post-treatment variables to improve the precision [11].

4.3.2 Fairness Evaluation

So far, we have examined the causal assumptions and their implications in the context of fairness evaluation. Once the plausibility of the assumptions are established, we can proceed to estimate the treatment effect of A on Y. While there are many approaches in estimating the causal effect,

we mainly focus on direct regression method. We first consider a case where post-treatment variables are absent. Under causal assumptions, treatment effect of A can be formulated as

$$\begin{aligned}\tau = \mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}\left[\mathbb{E}\left[Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}, A = 1\right] - \mathbb{E}\left[Y \mid \boldsymbol{X} = \boldsymbol{x}, A = 0\right]\right],\end{aligned}$$

which can be estimated from observational data via two separate regression models. When post-treatment variables are present, it may be helpful to use them in order to improve the precision of treatment effect estimates. In this case, simply conditioning on those variables will introduce bias in the analysis. Instead, we should treat them as dependants on A. In order to emphasize the causal effect of post-treatment variables on the potential outcomes, we consider potential outcomes $Y(A, \tilde{X}(A))$ that are indexed by both the treatment and the post-treatment counterfactuals. We estimate the treatment effect of A on Y is given as $\tau = \mathbb{E}\left[Y(1, \tilde{X}(1)) - Y(0, \tilde{X}(0))\right]$. In the mediation literature, this quantity is known as *total effect* [94].

Estimating the total effect poses a considerable identification challenge as it depends on four $(\tilde{X}(0), \tilde{X}(1), Y(0, \tilde{X}(0)), Y(1, \tilde{X}(1)))$ counterfactuals which are not simultaneously observed for any individual. To tackle this problem, we propose to use imputation [161] which is commonly used in causal inference literature to assign values to unobserved variables in the data. Precisely, in order to attain the causal effect of A on Y, we sequentially impute the missing variables were conditional on the previous step. Precisely, we first impute the counterfactual post-treatment variables $\tilde{X}$ as a function of the pre-treatment variables $\boldsymbol{X}$ and A. Next, we impute unobserved $Y(A, \tilde{X}(A))$ values as a function of pre-treatment variables $\boldsymbol{X}$, post-treatment counterfactuals $\tilde{X}$ and A. Similar sequential imputation techniques have been used in causal inference literature in order to evaluate the long-term impact of policy shifts [187].

4.3.3 Unfairness Mitigation

In the previous section, we focused solely on fairness evaluation which we formulated as a causal inference problem on the effect of A on Y. Here, we discuss how we can mitigate unfairness if the treatment effect of A on Y is non-zero. Similar to the previous section, we distinguish between pre- and post-treatment variables as the post-treatment variables are affected by A. The core idea of our unfairness mitigation approach is to adjust the post-treatment and outcome variables to achieve $\tau = 0$. The idea of adjusting downstream variables, affected by sensitive attributes, has been recently investigated in the fair ML literature and in the context for mitigating path-specific effects under the DAG framework [50]. In this work, we are interested in mitigating unfairness that attributed to a specific actor's decision-making process, rather than an entire causal path. Intuitively, our approach is based on the assumption that in a fair world, everyone is treated with no regard to their group membership. In other words, we deem a decision-making process fair if everyone is treated as if they belong to the same group, which we refer to as a baseline group. The baseline group can be viewed as either the majority group or a historically advantaged group.

We first consider a setting with no post-treatment variables and assume $\mathbb{E}[Y(1) - Y(0)] \neq 0$. Let $A = 0$ be the baseline group. If we had access to $Y(0)$ for every individual in the population, we could use that in order to learn a fair classifier. That is, if we observed the outcome of individuals had they belonged to the baseline group, we could use this data to learn a predictive model. The reason is that when we use $Y(0)$ to learn a model, we are effectively eliminating decision-maker's unfavorable attitude towards membership to group $A = 1$. Consequently, we can model the prediction problem as $\mathbb{P}(Y(0) = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. In the presence of post-treatment variables, we employ a similar approach in order to eliminate the discriminatory effects of A. Therefore, we formulate the prediction problem as $\mathbb{P}(Y(0) = 1 \mid \boldsymbol{X} = \boldsymbol{x}, \tilde{X}(0) = \tilde{x})$, in which we use $\tilde{X}(0)$ which is the value of the post-treatment variable had the individual belonged to group $A = 0$. Remarkably, under this formalization causal parity is automatically achieved as

$\mathbb{E}[Y(0) - Y(0)] = 0$. In words, we are practically assuming that an the potential outcomes for an individual is the same and is equal to $A = 0$, regardless of the observed value of A. A key challenge with this approach is that $Y(0)$ values are not observed for every individual. We leverage imputation from causal inference literature to tackle this problem [161].

4.4 Trade-offs under the Lens of Causality

We now turn to another important aspect of our analysis. We introduce causal variants of common statistical criteria of fairness to study their behavior under the causal lens.

4.4.1 Causal Fairness Definitions

We center our discussion on the criteria with known impossibility results in the fair ML literature.

Definition 4 (Conditional Causal Parity)**.** *A decision-making process achieves conditional causal parity if*

$$\mathbb{E}[Y(1) - Y(0) \,|\, \boldsymbol{X} = \boldsymbol{x}] = 0 \,\forall \boldsymbol{x} \in \mathcal{X}.$$

The above definition is closely related to conditional statistical parity which aims to evaluate fairness after controlling for a limited set of "legitimate" factors [101]. The set of legitimate factors significantly impacts the conclusions we draw. However, it is typically assumed as given, e.g., by domain experts. In contrast, in our definition $\boldsymbol{X}$ collects all the pre-treatment variables. Hence, once the nature of the intervention is explicitly defined, all remaining pre-treatment variables can be considered as legitimate since the main effect we aim to identify is the effect of the treatment.

Definition 5 (Causal Equalized Odd)**.** *A predictor $\hat{Y}$ satisfies causal equalized odds if:*

$$\mathbb{P}(\hat{Y} = 1 \mid Y(0) = 1) = \mathbb{P}(\hat{Y} = 1 \mid Y(1) = 1)$$

$$\mathbb{P}(\hat{Y} = 1 \mid Y(0) = 0) = \mathbb{P}(\hat{Y} = 1 \mid Y(1) = 0)$$

The above definition is the causal counterpart of equalized odds proposed in [85]. It states that the probability of receiving a positive prediction $\hat{Y} = 1$ in worlds where everyone is treated as $A = 0$ or $A = 1$ should be the same. Therefore, an individual does not have any preferences to be in either of these worlds since in either world the prediction is the same. Next, we define the causal variant of calibration [110]. Calibration is defined in the context of risk scores.

Definition 6 (Causal Calibration)**.** *Let $S \in \mathcal{S}$ denote a random variable encoding an individual's risk score. The risk assignment is well-calibrated within groups if it satisfies the following condition:*

$$\mathbb{P}(Y(0) = 1 \mid S = s) = \mathbb{P}(Y(1) = 1 \mid S = s)\ \forall s \in \mathcal{S}.$$

Causal calibration states that a risk score S should have the same meaning in worlds where everyone is treated as $A = 0$ or $A = 1$, i.e., the proportion of positive outcomes in either worlds should be the same for any fixed $S = s$. Subsequently, we can define causal positive predictive parity.

Definition 7 (Causal Positive Predictive Parity)**.** *A predictor $\hat{Y}$ satisfies causal positive predictive parity if:*

$$\mathbb{P}(Y(0) = 1 \mid \hat{Y} = 1) = \mathbb{P}(Y(1) = 1 \mid \hat{Y} = 1).$$

Causal predictive parity is applicable in the binary decision-making scenarios and has a similar interpretation as causal calibration in that it requires the proportion of positive outcomes in worlds with $A = 0$ and $A = 1$ to be the same for any fixed $\hat{Y} = 1$. Therefore, an individual with positive prediction does not feel being discriminated against since in both worlds, the rate of positive outcome is the same.

4.4.2 Trade-offs among Causal Criteria of Fairness

We investigate two main impossibility results known for the statistical fairness criteria and show that there is no fundamental disagreement between their causal variants.

Causal Parity and Conditional Causal Parity. It is easy to see that statistical parity and conditional statistical parity may not be satisfied simultaneously on a dataset. The Berkeley college admission study is a notorious example [40]. In this study, it was shown that while female students where admitted at a lower rate compared to the male students, after controlling for department choice, the difference in admission rates became insignificant among the two groups. This observation can be expressed formally as below.

Observation 1. *There exists a joint distribution* $p(\boldsymbol{X}, A, Y)$ *such that conditional statistical parity does not imply statistical parity, i.e.,* $\mathbb{E}[Y \mid \boldsymbol{X} = x, A = 1] - \mathbb{E}[Y \mid \boldsymbol{X} = x, A = 0] = 0 \; \forall \boldsymbol{x} \in \mathcal{X} \implies \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] = 0$.

Contrary to the above result, it is straightforward to show that conditional causal parity implies causal parity.

Proposition 10. *Conditional causal parity implies causal parity. Mathematically,*

$$\mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{X} = x] = 0 \; \forall \boldsymbol{x} \in \mathcal{X} \implies \mathbb{E}[Y(1) - Y(0)] = 0.$$

Proof. The proof follows simply from taking the expectation over $\boldsymbol{X}$. ■

The intuition behind the above result is that $\mathbb{E}[Y \mid A]$ merely measures the statistical dependence between Y and A and does not differentiate between different sources of dependence, e.g.,

female students applying for more competitive departments than male students, or a discriminatory admission process. We note that conditional causal parity is a more stringent requirement than causal parity and the reverse implication does not generally hold true in Proposition 10.

Causal Positive Predictive Parity and Causal Equalized Odds. It is well-known that one can not achieve positive predictive parity or calibration together with equalized odds simultaneously unless either the base rates $\mathbb{P}(Y \mid A = a)$ are equal or the classifier is perfect [110, 51]. Here, we show no such restrictions are necessary for their causal variants. We first define $f : \mathcal{S} \to \{0, 1\}$ as a mapping from the risk score S to binary prediction $\hat{Y}$. For example, $f(S) = \mathbb{I}(S > \theta)$ classifying the data points based on a threshold.

We now present our main results.

Theorem 2. *Causal calibration implies causal parity and causal equalized odds.*

Proof. First, we note that causal calibration implies causal parity by taking the expectation over $s \in \mathcal{S}$. Also, we can show $\forall s \in \mathcal{S}$, the equality $\mathbb{P}(Y(0) = 1 \mid S = s) = \mathbb{P}(Y(1) = 1 \mid S = s)$ implies:

$$\begin{aligned}
&\mathbb{P}(S = s \mid Y(0) = 1)\mathbb{P}(Y(0) = 1) = \mathbb{P}(S = s \mid Y(1) = 1)\mathbb{P}(Y(1) = 1) \\
\Rightarrow &\mathbb{P}(S = s \mid Y(0) = 1) = \mathbb{P}(S = s \mid Y(1) = 1) \\
\Rightarrow &\mathbb{P}(\hat{Y} = f(s) \mid Y(0) = 1) = \mathbb{P}(\hat{Y} = f(s) \mid Y(1) = 1).
\end{aligned}$$

Similarly, we can show:

$$\mathbb{P}(Y(0) = 0 \mid S = s) = \mathbb{P}(Y(1) = 0 \mid S = s) \Rightarrow \mathbb{P}(\hat{Y} = f(s) \mid Y(0) = 0) = \mathbb{P}(\hat{Y} = f(s) \mid Y(1) = 0).$$

■

Causal parity, i.e., $P(Y(0) = 1) = P(Y(1) = 1)$, is satisfied if decisions are made regardless of one's group membership and is different from the equal base rate assumption which does not necessarily hold in many applications. The above result shows that there is an inherent compatibility between different causal fairness criteria as achieving one automatically implies one or two other criteria.

Lemma 4. *If $A/B = C/D$ and $(1 - A)/(1 - B) = (1 - C)/(1 - D)$, then $A = C$ and $B = D$.*

Proof.

$$\left.\begin{aligned} \frac{A}{B} = \frac{C}{D} \Rightarrow \frac{A-B}{B} = \frac{C-D}{D} \\ \frac{1-A}{1-B} = \frac{1-C}{1-D} \Rightarrow \frac{A-B}{1-B} = \frac{C-D}{1-D} \end{aligned}\right\} \Rightarrow \frac{1-B}{B} = \frac{1-D}{D}.$$

It follows that $A = C$ and $B = D$. ■

Theorem 3. *Causal equalized odds implies causal parity and causal positive predictive parity.*

Proof.

$$\mathbb{P}(\hat{Y} = 1 \mid Y(0) = 1) = \mathbb{P}(\hat{Y} = 1 \mid Y(1) = 1) \Rightarrow \frac{\mathbb{P}(Y(0) = 1 \mid \hat{Y} = 1)}{\mathbb{P}(Y(0) = 1)} = \frac{\mathbb{P}(Y(1) = 1 \mid \hat{Y} = 1)}{\mathbb{P}(Y(1) = 1)}.$$
$$\mathbb{P}(\hat{Y} = 1 \mid Y(0) = 0) = \mathbb{P}(\hat{Y} = 1 \mid Y(1) = 0) \Rightarrow \frac{\mathbb{P}(Y(0) = 0 \mid \hat{Y} = 1)}{\mathbb{P}(Y(0) = 0)} = \frac{\mathbb{P}(Y(1) = 0 \mid \hat{Y} = 1)}{\mathbb{P}(Y(1) = 0)}.$$

From Lemma 4, it follows that $\mathbb{P}(Y(0) = 1) = \mathbb{P}(Y(1) = 1)$ and $\mathbb{P}(Y(0) = 1 \mid \hat{Y} = 1) = \mathbb{P}(Y(1) = 1 \mid \hat{Y} = 1)$, where the first and second equations correspond to causal parity and causal positive predictive parity, respectively. ■

We conclude this section by providing a complementary result that relates conditional causal parity to causal calibration and causal positive predictive parity.

Proposition 11. *Given a risk score as a function of pre-treatment variables $\boldsymbol{X}$, i.e., $S = h(\boldsymbol{X})$, it holds that conditional causal parity implies causal calibration and causal positive predictive parity.*

$$
\begin{aligned}
&\mathbb{P}(Y(0) = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Y(1) = 1 \mid \boldsymbol{X} = \boldsymbol{x})\ \forall \boldsymbol{x} \in \mathcal{X} \\
\Rightarrow &\mathbb{P}(Y(0) = 1 \mid \boldsymbol{X} \in h^{-1}(s)) = \mathbb{P}(Y(1) = 1 \mid \boldsymbol{X} \in h^{-1}(s))\ \forall s \in \mathcal{S} \\
\Rightarrow &\mathbb{P}(Y(0) = 1 \mid h(\boldsymbol{X}) = s) = \mathbb{P}(Y(1) = 1 \mid h(\boldsymbol{X}) = s)\ \forall s \in \mathcal{S} \\
\Rightarrow &\mathbb{P}(Y(0) = 1 \mid \hat{Y} = f(s)) = \mathbb{P}(Y(1) = 1 \mid \hat{Y} = f(s))\ \forall f(s) \in \{0, 1\}.
\end{aligned}
$$

Consequently, causal parity and causal equalized odds will be satisfied. The above result implies that for a given set of pre-treatment variables $\boldsymbol{X}$, if conditional causal parity is satisfied, all other causal fairness criteria discussed in the present work will be satisfied provided that the risk score function h and the classifier f are a functions of the pre-treatment variables. Conditional causal parity can be achieved using the imputation technique described in the previous section. Finally, it is important to note that the above results are based on the assumption that the joint distribution of variables is known. In practice, factors such as inadequate sample sizes, modelling choices, hyper-parameter selection, etc. can influence the performance of models across different groups. In the standard ML setting, previous work has aimed to address some of these limitations through careful model selection or additional training data collection, etc. [49].

4.5 Computational Results

We consider a stylized hiring scenario to illustrate our causal unfairness evaluation and mitigation approach. Specifically, we consider a decision-making process that involves two interactions:

interview and final hiring decision. We study how the timing of the intervention impacts our conclusions.

We use A to represent gender, which we draw from a Bernoulli distribution $Bern(0.75)$ with the majority class being male $A = 1$. An individual's qualification is described by a random variable X drawn from a normal distribution $\mathcal{N}(2\alpha(A-0.5), 1)$, where α controls the difference in the average qualifications between genders. Each candidate has a score S reflecting their performance during the interview. We model the score as a binary variable whose mean depends on the qualifications and possibly gender. We have $\mathbb{P}(S = 1) = \sigma\left(X + 2\beta(A - 0.5)\right)$, where $\sigma(z) = 1/(1 + e^{-z})$ is the logistic function and $\beta \geq 0$ determines the level of discrimination in S, e.g., when $\beta > 0$ being a male $A = 1$ increases one's probability of receiving a higher score. Subsequently, a decision Y is made indicating whether the candidate receives an offer or not. We use the probabilistic model $\mathbb{P}(Y = 1) = \sigma\left(X + S + 2\gamma(A - 0.5)\right)$, with $\gamma \geq 0$ controlling the level of discrimination in Y for a fixed X, S. The vector of potential outcomes, for both S and Y, can be obtained by substituting the respective value of A in the model. We present results across a wide range of α, β and γ values.

According to our causal framework, we need to specify the point in time from which the effect of gender needs to be assessed. There are two possibilities: after interview is conducted or before the interview (as one may be concerned about an unfair interview process). Naturally, the above choice will impact our conclusions about whether the system is fair or not. We generate 100,000 data points $(X, A, S(0), S(1), Y(0, S(0)), Y(1, S(1)))$ according to the process explained above. For post-interview fairness evaluation we can use the observed S values as the score is a pre-treatment variable. However, when evaluating fairness before the interview, the score becomes post-treatment. In order to impute the missing counterfactual score values $S(A)$, we use logistic regression to model $\mathbb{P}(S = 1 \mid \boldsymbol{X} = \boldsymbol{x}, A = a),\ a \in \{0, 1\}$, from which we sample (10 samples). We use a second logistic regression $\mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, S(a) = s, A = a),\ \forall a, s \in \{0, 1\}$ to impute $Y(A, S(A))$ values. This approach is based on multiple imputation in the causal inference

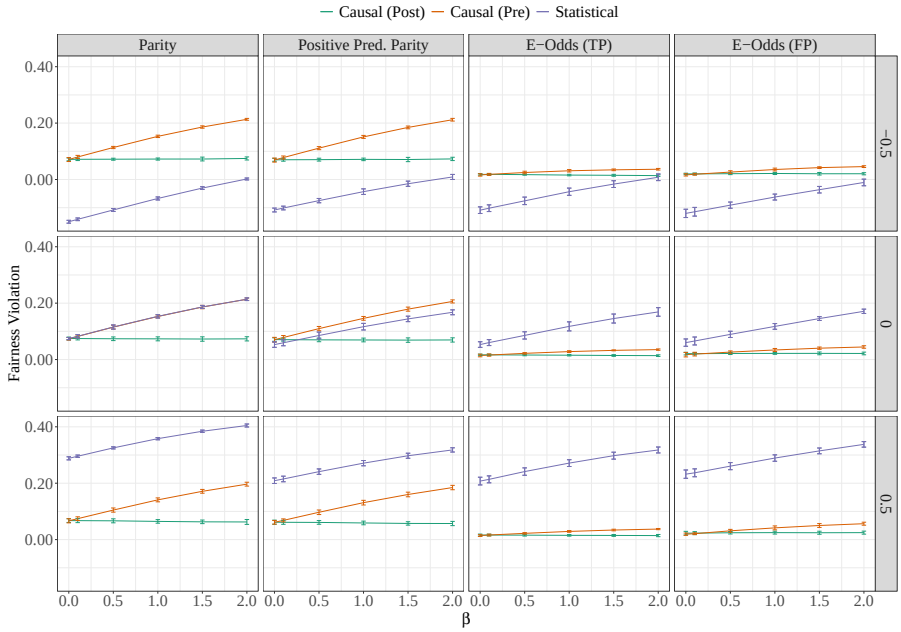


Figure 4.2: Synthetic results in the hiring scenario. Colors denote the evaluation method: causal pre-interview, causal post-interview and statistical. From top to bottom, each row corresponds to a different value of $\alpha \in \{-0.5, 0, 0.5\}$. Column are different fairness evaluation criteria. On the $x-$axis, we vary the value of β, which reflects the dependence of the interview score on one's gender. The $y-$axis shows fairness violation across four different definitions. We note that for causal approaches we use the causal variants of the fairness definitions. The value of γ is set to 0.2. The error bars show 95% confidence interval. Depending on the joint setting of the parameters, statistical criteria may erroneously result in an over- or under-estimation of fairness violation. Further, post-interview fairness evaluation does not capture discrimination at earlier points in time.

literature [161]. We then use these counterfactual values in the expression that evaluates the treatment effect of A.

We compare our causal criteria against statistical fairness definitions, where we measure the fairness violation of a logistic regression model trained to predict Y using observed values of X, S and A. Figure 4.2 depicts a summary of results. We can make several key observations. First, the post-interview causal plot remains flat across different values of α, β exhibiting a constant fairness violation at 0.07 due to constant $\gamma = 0.2$ which is independent of prior discrimination in the interview step. This suggests that early discrimination can not be captured when one chooses a later time as the point of fairness investigation. In other words, any unfairness in the pre-treatment variables used in the analysis will remain undetected. Pre- and post-interview lines

only intersect at $\beta = 0$ and pre-interview fairness violation increases monotonically with β across all causal fairness definitions. Statistical fairness definitions exhibit significantly different results. For example, when $\alpha = -0.5$, all statistical lines lie below the causal ones which suggests that they underestimate the true level of discrimination. This is due to the fact that when $\alpha < 0$, males qualification is lower than females on average. However, since $\beta, \gamma > 0$ the interview score and the final decision are in favor of male candidates. Since the statistical criteria fail to disentangle different sources of disparities, these opposing effects are cancelled, resulting in lower estimates of unfairness. On the other hand, when $\alpha > 0$, these effects reinforce each other resulting in an over-estimation of unfairness. Only when $\alpha = 0$, do statistical parity and causal parity, in Causal (Pre), match which indicates the sensitive of statistical criteria to baseline differences between groups (average qualifications). For $\beta, \gamma = 0$ (no discrimination in interview or the hiring process), our results indicate near-zero estimates for all causal definitions of fairness across different values of α. This confirms that it is indeed possible to satisfy different causal fairness definitions simultaneously, even when there are baseline differences between the qualifications of different groups. Conversely, statistical criteria yield non-zero estimates except for the case where $\alpha, \beta, \gamma = 0$ which points to the equal base rate condition highlighted in previous work [110].

Next, we study the power of our approach to mitigate unfairness. We focus on the setting where $\alpha = 0$ and $\beta, \gamma \neq 0$. This is because we aim to remove unfairness in the decision-making process, which is associated with β and γ. Since statistical approaches are not able to disentangle different sources of unfairness, by setting α we are able to compare our results against those criteria. Specifically, we train a model using our approach by imputing the missing potential outcomes. We compare the accuracy and fairness violation with an unconstrained model (No Fairness), as well as the same model after applying one of three common unfairness mitigation algorithms in the literature: pre-processing method (Re-weighting) in [99] which generates weights for the training examples in each combination of A and Y differently to ensure statistical parity, in-processing method (PrejudiceRemover) of [102] that adds a regularization term to the learning

	Fairness Violation (Statistical Criteria)				
	Parity	Positive Pred. Parity	E-Odds (TP)	E-Odds (FP)	Acc. (%)
No Fairness	0.31	0.02	0.26	0.21	72.7
Re-weighting	0.10	0.11	0.04	0.03	71.8
PrejudiceRemover	0.16	0.04	0.02	0.24	74.0
RejectOption	0.05	0.19	0.09	0.16	72.0
Causal (Pre)	0.03	0.14	-0.02	-0.04	70.0
Causal (Post)	0.17	0.07	-0.11	-0.09	72.0

Table 4.1: Fairness violation of statistical criteria and the classification accuracy.

objective, and post-processing approach (RejectOption) in [100] which gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary. We rely on the implementations in AI Fairness 360 package [22]. The training model used in all of the methods is a logistic regression model. For fairness violation, we consider both statistical criteria and their causal variants.

Table 4.1 summarizes the statistical fairness violation results for $\beta = 1.0$ and $\gamma = 0.2$. Among all fair baselines, RejectOption and Causal (Pre) perform significantly better in terms of statistical parity. Despite the fact that other fair baselines are also designed to remove average disparities between groups, they still exhibit significant disparities. On the other hand, RejectOption performs worse than Causal (Pre) with respect to all other criteria. We also note the difference in pre- and post-interview results. The increase in parity violation in Causal (Post) can be explained by the fact that it only adjusts for the outcome variable and assumes disparities in the interview score are acceptable. Disparities in score will in turn result in different outcomes across groups but in post-interview analysis this effect remains undetected. Finally, in terms of accuracy, Causal (Post) achieves comparable results to the other methods. The difference in the accuracy of Causal (Pre) and (Post) is in part due to the fact that the accuracy is measured with respect to the observed unfair outcomes. As a result, Causal (Pre) which significantly reduces the gap between female and male candidates may not conform to the historical decisions. Finally, these results highlight the importance of determining the timing of the intervention. Specifically, they suggest that through the causal framework, we are able to identify and remove sources of

	Fairness Violation (Causal Criteria)			
	Parity	Positive Pred. Parity	E-Odds (TP)	E-Odds (FP)
Causal (Pre)	0.00	0.00	0.00	0.00
Causal (Post)	0.06	0.02	0.05	0.01

Table 4.2: Fairness violation of causal criteria.

disparities by actively adjusting the affected variables. Finally, we evaluated our approach based on causal criteria of fairness, choosing pre-interview as the starting time of fairness assessment. In Table 4.2, we observe no violation of fairness in Causal (Pre) as expected. However, Causal (post) exhibits small violations which is due to the fact that it only mitigates unfairness due to γ and in the hiring decision.

4.6 Conclusion and Broader Impact

As empirical evidence on ethical implications of algorithmic decision-making is growing [159, 131, 138, 168], a variety of approaches have been proposed to evaluate and minimize the harms of these algorithms. In the statistical fairness literature, it is well-established that it is not possible to satisfy every fairness criterion simultaneously, which results in significant trade-offs in selecting a metric. On the other hand, in the causal fairness literature, there is substantial ambiguity around how the proposed methods should be applied to a particular problem. Also, these methods rely on assumptions that are often too strong to be applicable in practice. In this work, we addressed some of these limitations.

First, we illustrated the utility of applying concepts from the "potential outcomes framework" to algorithmic fairness problems. In particular, we emphasized the timing and nature of the intervention as two key aspects of causal fairness analysis. That is, for any valid causal analysis, it is critical to precisely define the starting point of the fairness evaluation and the postulated intervention. We argue that fairness evaluation is not a static problem and unfairness can happen at various points and within and across multiple domains. This is contrast with methods that

rely on fixed DAG models. Next, we demonstrated how such a causal framework can address the limitations of existing approaches. Specifically, our theoretical investigation indicates that there is an inherent compatibility between the causal fairness definitions we propose. Finally, we showed the effectiveness of our approach in evaluating and mitigating unfairness associated with different stages of decision-making. We hope that our empirical observations spark additional work on collecting new datasets that lend themselves to temporal fairness evaluation.

Conclusion and Future Work

This thesis identifies and addresses several challenges with respect to designing fair algorithmic social interventions. The contributions of this thesis are both technical and practical. On a technical level, this work presents novel computational models that capture real-world complexities such as data and resource scarcity as well as fairness considerations. Specifically, this thesis investigates the interplay of fairness with data limitation, data biases and resource scarcity to develop fair and efficient algorithmic frameworks that solve the resulting optimization models. From a practical perspective, it contributes different intervention models to prevention and social sciences. In particular, this work proposes to use social network information to inform gatekeeper training for suicide prevention and enhance community resilience against natural hazards. It also presents the first use of quantitative techniques to inform these interventions. Finally, this work proposes an implementable policy model to allocate scarce housing resources to individuals experiencing homelessness.

All in all, this thesis covers a subset of challenges in developing effective algorithmic social interventions and much work remains to be done. Specifically, due to the socio-technical nature of fairness research, there are several social and legal considerations before any of these solutions can be deployed. For instance, one question is related to the choice of social groups in the group-based definitions of fairness. This problem is known as *intersectionality* fairness which states that disparities can be amplified in subgroups that combine membership from different categories

(e.g., race and gender), especially if such a subgroup is particularly under-represented historically [52, 79]. While the methodologies presented in this work cater for fairness over intersection of different groups, there is still ambiguity in how these groups should be identified. Individual-based definitions, on the other hand, are often too restrictive and are not readily applicable to problems that suffer from resource scarcity as we can not ensure every individual receives a fair share of resources.

There are also issues with respect to legal compatibility of these solutions. In particular, U.S. law prohibits policies that differentiate between individuals based on protected attributes such as race, gender or age [186]. On the other hand, there are exceptions that allow policies that aim to overcome present disparities of past practices, policies, or barriers by prioritizing resources for underserved or vulnerable groups. Existing methods typically use the information about one's group membership to ensure fair distribution of intervention benefits which may not be immediately compatible with the legal frameworks. As a result, further research is needed to ensure the usability of these solutions from the legal perspectives.

There are also issues related to unobserved confounders. When estimating the treatment effect of different interventions, a common assumption is that all the confounding factors between treatment and outcome are captured in the data. In observational studies, this assumption is practically impossible to verify and there is always a threat to its validity. As a result, one avenue of research could consider robustifying the estimates against such unobserved factors.

Furthermore, due to the context-dependent nature of fairness, re-purposing algorithmic solutions designed for one social context may be misleading or even inaccurate when applied to a different context. Existing frameworks of fairness also suffer from a lack of expressiveness, i.e., they provide point-solutions tailored to a specific context. This thesis aimed to tackle this problem by presenting two unifying frameworks for fairness: fairness in social-network interventions and a causal framework for fairness for decision-making under observational data. Even though the presented frameworks encompass a wide-range of problems, they are not universal. For instance,

social network-based intervention models with non-submodular utility functions are not handled by the presented framework and further research is necessary to tackle those problems.

Bibliography

[1] Ignacio Abasolo and Aki Tsuchiya. Exploring social welfare functions and violation of monotonicity: an example from inequalities in health. *Journal of Health Economics*, 23(2):313–329, 2004.

[2] Ivo Adan and Gideon Weiss. A skill based parallel service system under FCFS-ALIS — steady state, overloads, and abandonments. *Stochastic Systems*, 4(1):250–299, 2014.

[3] Philipp Afèche, René Caldentey, and Varun Gupta. On the Optimal Design of a Bipartite Matching Queueing System. *Operations Research*, 2021.

[4] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[5] Faez Ahmed, John P. Dickerson, and Mark Fuge. Diverse weighted bipartite b-matching. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 35–41. AAAI Press, 2017.

[6] Junaid Ali, Mahmoudreza Babaei, Abhijnan Chakraborty, Baharan Mirzasoleiman, Krishna Gummadi, and Adish Singla. On the fairness of time-critical influence maximization in social networks. *arXiv*, abs/1905.06618, 2019.

[7] Kareem Amin, Satyen Kale, Gerald Tesauro, and Deepak Turaga. Budgeted prediction with expert advice. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Propublica*, 2016.

[9] Bryan Anselm. Suicide on campus. *The New York Times*, August 25 2015.

[10] Barics Ata and Mustafa H. Tongarlak. On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems*, 74(1):65–104, 2013.

[11] Susan Athey, Raj Chetty, Guido W. Imbens, and Hyunseung Kang. The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely. *SSRN*, 2019.

[12] LOS ANGELES HOMELESS SERVICES AUTHORITY. Report and recommendations of the ad hoc committee on black people experiencing homelessness. Technical report, Feb 2019.

[13] Mohammad-Javad Azizi, Phebe Vayanos, Bryan Wilder, Eric Rice, and Milind Tambe. Designing fair, efficient, and interpretable policies for prioritizing homeless youth for housing resources. In *International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 35–51. Springer, 2018.

[14] Chaithanya Bandi, Nikolaos Trichakis, and Phebe Vayanos. Robust multiclass queuing theory for wait time estimation in resource allocation systems. *Management Science*, 65(1):152–187, 2018.

[15] Siddharth Barman, Arpita Biswas, Sanath Krishnamurthy, and Yadati Narahari. Groupwise maximin fair allocation of indivisible goods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[16] Siddharth Barman, Sanath Kumar Krishnamurthy, and Rohit Vaish. Finding fair and efficient allocations. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 557–574, 2018.

[17] Anamika Barman-Adhikari, Stephanie Begun, Eric Rice, Amanda Yoshioka-Maxwell, and Andrea Perez-Portillo. Sociometric network structure and its association with methamphetamine use norms among homeless youth. *Social science research*, 58:292–308, 2016.

[18] Mohammad-Hossein Bateni, Yiwei Chen, Dragos F. Ciocan, and Vahab Mirrokni. Fair resource allocation in a volatile marketplace. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 819–819. ACM, 2016.

[19] MohammadHossein Bateni, Yiwei Chen, Dragos Florin Ciocan, and Vahab Mirrokni. Fair resource allocation in a volatile marketplace. *Operations Research*, 70(1):288–308, 2022.

[20] Xiaohui Bei, Zihao Li, Jinyan Liu, Shengxin Liu, and Xinhang Lu. Fair division of mixed divisible and indivisible goods. *Artificial Intelligence*, 293:103436, 2021.

[21] Xiaohui Bei, Xinhang Lu, Pasin Manurangsi, and Warut Suksompong. The price of fairness for indivisible goods. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 81–87. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[22] Rachel K.E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv*, 2018.

[23] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[24] Nawal Benabbou, Mithun Chakraborty, Xuan-Vinh Ho, Jakub Sliwinski, and Yair Zick. Diversity constraints in public housing allocation. In *Proceedings of the 17th International*

Conference on Autonomous Agents and MultiAgent Systems, pages 973–981. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[25] Jacques F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Computational Management Science*, 2(1):3–19, 2005.

[26] Abram Bergson. A reformulation of certain aspects of welfare economics. *The Quarterly Journal of Economics*, 52(2):310–334, 1938.

[27] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[28] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[29] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4), 2004.

[30] Dimitris Bertsimas and Constantine Caramanis. Finite adaptability in multistage linear optimization. *IEEE Transactions on Automatic Control*, 55(12):2751–2766, 2010.

[31] Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru. Optimal Prescriptive Trees. *INFORMS Journal on Optimization*, 1(2):164–183, 2019.

[32] Dimitris Bertsimas and Iain Dunning. Multistage robust mixed-integer optimization with adaptive partitions. *Operations Research*, 64(4):980–998, 2016.

[33] Dimitris Bertsimas, Vivek Farias, and Nikolaos Trichakis. The price of fairness. *Operations Research*, 59(1):17–31, 2011.

[34] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Operations Research*, 61(1):73–87, 2013.

[35] Dimitris Bertsimas and Angelos Georghiou. Design of near optimal decision rules in multistage adaptive mixed-integer optimization. *Operations Research*, 63(3):610–627, 2015.

[36] Dimitris Bertsimas and Angelos Georghiou. Binary decision rules for multistage adaptive mixed-integer optimization. *Mathematical Programming*, 167(2):395–433, 2018.

[37] Dimitris Bertsimas and John Tsitsiklis. Introduction to linear programming. *Athena Scientific*, 1:997, 1997.

[38] Dimitris Bertsimas and Phebe Vayanos. Data-driven learning in dynamic pricing using adaptive optimization. *Optimization Online*, 2017.

[39] Dimitris Bertsimas and Robert Weismantel. *Optimization over integers*, volume 13. 2005.

[40] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 1975.

[41] Arpita Biswas and Siddharth Barman. Fair division under cardinality constraints. In *IJCAI*, pages 91–97, 2018.

[42] Ilija Bogunovic, Slobodan Mitrović, Jonathan Scarlett, and Volkan Cevher. Robust submodular maximization: A non-uniform partitioning approach. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 508–516. JMLR. org, 2017.

[43] Thomas Bonald and Laurent Massoulié. Impact of fairness on internet performance. In *Joint International Conference on Measurements and Modeling of Computer Systems*, pages 82–91, 2001.

[44] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, 2009.

[45] Francisco Castro, Hamid Nazerzadeh, and Chiwei Yan. Matching queues with reneging: a product form solution. *Queueing Systems*, 96(3-4):359–385, 2020.

[46] Hau Chan, Eric Rice, Phebe Vayanos, Milind Tambe, and Matthew Morton. Evidence from the past: AI decision AIDS to improve housing systems for homeless youth. In *AAAI Fall Symposium - Technical Report*, volume FS-17-01 - FS-17-05, Stanford University, United States, 2017. AAAI Press.

[47] André Chassein, Marc Goerigk, Jannis Kurtz, and Michael Poss. Faster algorithms for min-max-min robustness for combinatorial problems with budgeted uncertainty. *European Journal of Operational Research*, 2019.

[48] Chandra Chekuri, Jan Vondrak, and Rico Zenklusen. Dependent randomized rounding via exchange properties of combinatorial structures. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 575–584. IEEE, 2010.

[49] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, volume 2018-December, 2018.

[50] Silvia Chiappa. Path-specific counterfactual fairness. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2019.

[51] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *3rd Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[52] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[53] Vincent Conitzer, Rupert Freeman, Nisarg Shah, and Jennifer W. Vaughan. Group fairness for the allocation of indivisible goods. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[54] 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*, pages 797–806, 2017.

[55] Koby Crammer, Jaz Kandola, and Yoram Singer. Online classification on a budget. *Advances in neural information processing systems*, 16, 2003.

[56] Sergio Currarini, Matthew O. Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

[57] Hugh Dalton. The Measurement of the Inequality of Incomes. *The Economic Journal*, 30(119):348–361, 1920.

[58] Department of Housing and Urban Development. Restoring Affirmatively Furthering Fair Housing Definitions and Certifications. Technical report, Office of Fair Housing and Equal Opportunity, HUD., 2021.

[59] John P. Dickerson and Tuomas Sandholm. FutureMatch: Combining human value judgments and machine learning to match in dynamic environments. In *Proceedings of the National Conference on Artificial Intelligence*, volume 1, pages 622–628, Austin, Texas, United States, 2015. AAAI press.

[60] Yichuan Ding, S. Thomas McCormick, and Mahesh Nagarajan. A fluid model for one-sided bipartite matching queues with match-dependent rewards. *Operations Research*, 69(4), 2021.

[61] Kate Donahue and Jon Kleinberg. Fairness and utilization in allocating resources with uncertain demand. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 658–668, 2020.

[62] Miroslav Dudik, John Langford, and Hong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 1097–1104, Bellevue Washington USA, 2011. Omnipress 2600 Anderson St Madison WI United States.

[63] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *3rd Innovations in Theoretical Computer Science*, pages 214–226, 2012.

[64] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 170–179, 2019.

[65] Paul Erd6s. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[66] Brandon Fain, Kamesh Munagala, and Nisarg Shah. Fair allocation of indivisible public goods. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 575–592, 2018.

[67] Mohammad M. Fazel-Zarandi and Edward H. Kaplan. Approximating the first-come, first-served stochastic matching model with Ohm's law. *Operations Research*, 66(5):1423–1432, 2018.

[68] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

[69] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2015-August, 2015.

[70] Stephen E. Fienberg and Stanley S. Wasserman. Categorical data analysis of single sociometric relations. *Sociological methodology*, 12:156–192, 1981.

[71] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 489–503, 2021.

[72] Benjamin Fish, Ashkan Bashardoust, Danah Boyd, Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Gaps in information access in social networks? In *The World Wide Web Conference*, pages 480–490. ACM, 2019.

[73] Duncan Karl Foley. *Resource allocation and the public sector*. Yale University, 1966.

[74] Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *arXiv*, abs/1609.07236, 2016.

[75] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[76] Anthony Fulginiti, Aida Rahmattalabi, Jarrod Call, Phebe Vayanos, and Eric Rice. Using algorithmic solutions to address gatekeeper training issues for suicide prevention on college campuses. *Artificial Intelligence for Healthcare: Interdisciplinary Partnerships for Analytics-driven Improvements in a Post-COVID World*, page 83, 2022.

[77] Anthony Fulginiti, Aida Rahmattalabi, Jarrod Call, Phebe Vayanos, and Eric Rice. Using algorithmic solutions to address gatekeeper training issues for suicide prevention on

college campuses. *Artificial Intelligence for Healthcare: Interdisciplinary Partnerships for Analytics-driven Improvements in a Post-COVID World*, page 83, 2022.

[78] Vincent A. Fusaro, Helen G. Levy, and H. Luke Shaefer. Racial and Ethnic Disparities in the Lifetime Prevalence of Homelessness in the United States. *Demography*, 55(6):2119–2128, 2018.

[79] Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR, 2021.

[80] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[81] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[82] Steven N. Goodman, Sharad Goel, and Mark R. Cullen. Machine learning, health disparities, and causal reasoning, 2018.

[83] D. James Greiner and Donald B. Rubin. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3), 2011.

[84] Grani A. Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann. K-adaptability in two-stage robust binary programming. *Operations Research*, 63(4):877–891, 2015.

[85] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[86] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in Neural Information Processing Systems*, 31, 2018.

[87] Meghan Henry, Tanya de Sousa, Caroline Roddey, Swati Gayen, Thomas Joe Bednar, and Abt Associates. AHAR: Part 1—PIT Estimates of Homelessness in the US HUD Exchange. Technical report, The U.S. Department of Housing and Urban Development, Office of Community Planning and Development, 2020.

[88] Miguel a Hernán and James M Robins. Causal Inference Book. *Http://Www.Hsph.Harvard.Edu/Miguel-Hernan/Causal-Inference-Book/*, 2013.

[89] C Hill, H Hsu, M Holguin, M Morton, H Winetrobe, and E Rice. An examination of housing interventions among youth experiencing homelessness: an investigation into racial/ethnic and sexual minority status. *Journal of Public Health*, 2021.

[90] Christopher Hitchcock and Judea Pearl. Causality: Models, Reasoning and Inference. *The Philosophical Review*, 110(4), 2001.

[91] Adam G Horwitz, Taylor McGuire, Danielle R Busby, Daniel Eisenberg, Kai Zheng, Jacqueline Pistorello, Ronald Albucher, William Coryell, and Cheryl A King. Sociodemographic differences in barriers to mental health care among college students at elevated suicide risk. *Journal of affective disorders*, 271:123–130, 2020.

[92] Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

[93] Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, 2020.

[94] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1), 2010.

[95] Guido W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics, 2020.

[96] Michael Isaac, Brenda Elias, Laurence Y. Katz, Shay-Lee Belik, Frank P. Deane, Murray W. Enns, Jitender Sareen, and Swampy Cree Suicide Prevention Team (12 members). Gatekeeper training as a preventative intervention for suicide: a systematic review. *The Canadian Journal of Psychiatry*, 54(4):260–268, 2009.

[97] Maxwell Izenberg, Ryan Brown, Cora Siebert, Ron Heinz, Aida Rahmattalabi, and Phebe Vayanos. A community-partnered approach to social network data collection for a large and partial network. *Field Methods*, page 1525822X221074769, 2022.

[98] Nathanael Jo, Sina Aghaei, Andres Gomez, and Phebe Vayanos. Learning Optimal Prescriptive Trees from Observational Data. 2021.

[99] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 2012.

[100] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2012.

[101] Faisal Kamiran, Indre Žliobaite, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3), 2013.

[102] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7524 LNAI, 2012.

[103] Edward Harris Kaplan. *Managing the Demand for Public Housing.* PhD thesis, MIT, 1984.

[104] Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[105] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[106] Moniba Keymanesh, Tanya Berger-Wolf, Micha Elsner, and Srinivasan Parthasarathy. Fairness-aware Summarization for Justified Decision-Making, 7 2021.

[107] Aria Khademi, David Foley, Sanghack Lee, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2019.

[108] Shakeer Khan and Elie Tamer. Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica*, 78(6):2021–2042, 2010.

[109] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, volume 2017-December, 2017.

[110] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Conference on Innovations in Theoretical Computer Science*, pages 43:1–43:23, 2017.

[111] Jon Kleinberg, Yuval Rabani, and Éva Tardos. Fairness in routing and load balancing. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 568–578. IEEE, 1999.

[112] Issa Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review*, 113(5), 2019.

[113] Andreas Krause and Carlos Guestrin. Near-optimal observation selection using submodular functions. In *AAAI*, volume 7, pages 1650–1654, 2007.

[114] Amanda Kube, Sanmay Das, and Patrick J. Fowler. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 622–629, Honolulu, Hawaii, United States, 2019. AAAI Press.

[115] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 2017-December, 2017.

[116] Maria Kyropoulou, Warut Suksompong, and Alexandros A Voudouris. Almost envy-freeness in group resource allocation. *Theoretical Computer Science*, 841:110–123, 2020.

[117] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *49th Annual Meeting of the Association for Computational Linguistics*, pages 510–520, 2011.

[118] Cindy H Liu, Courtney Stevens, Sylvia HM Wong, Miwa Yasui, and Justin A Chen. The prevalence and predictors of mental health diagnoses and suicide among us college students: Implications for addressing disparities in service use. *Depression and anxiety*, 36(1):8–17, 2019.

[119] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 2019.

[120] Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1), 2018.

[121] Avishai Mandelbaum and Alexander L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research*, 52(6):836–855, 2004.

[122] Alexandre Marcellesi. Is race a cause? *Philosophy of Science*, 80(5), 2013.

[123] Michael Marmot, Sharon Friel, Ruth Bell, Tanja AJ Houweling, Sebastian Taylor, Commission on Social Determinants of Health, et al. Closing the gap in a generation: health equity through action on the social determinants of health. *The lancet*, 372(9650):1661–1669, 2008.

[124] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[125] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[126] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[127] Norweeta Milburn, Earl Edwards, Dean Obermark, and Janey Rountree. Inequity in the Permanent Supportive Housing System in Los Angeles: Scale, Scope and Reasons for Black Residents' Returns to Homelessness. Technical report, California Policy Lab, 2021.

[128] Clair Miller. Can an algorithm hire better than a human? *The New York Times*, June 25 2015. Ret. 4/28/2016.

[129] Charles E. Mitchell. An analysis of the U.S. Supreme Court's decision in Ricci v. DeStefano: The New Haven firefighter's case. *Public Personnel Management*, 42(1), 2013.

[130] Kaname Miyagishima. Fair criteria for social decisions under uncertainty. *Journal of Mathematical Economics*, 80:77–87, 2019.

[131] John Monahan and Jennifer L. Skeem. Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology*, 12, 2016.

[132] Jacob M. Montgomery, Brendan Nyhan, and Michelle Torres. How Conditioning on Post-treatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science*, 62(3), 2018.

[133] Matthew H. Morton, Amy Dworsky, Jennifer L. Matjasko, Susanna R. Curry, David Schlueter, Raúl Chávez, and Anne F. Farrell. Prevalence and Correlates of Youth Homelessness in the United States. *Journal of Adolescent Health*, 62(1), 2018.

[134] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018.

[135] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 2, pages 6–2, 2018.

[136] George L Nemhauser and Laurence A Wolsey. Maximizing submodular set functions: formulations and analysis of algorithms. In *North-Holland Mathematics Studies*, volume 59, pages 279–301. Elsevier, 1981.

[137] Quan Nguyen, Sanmay Das, and Roman Garnett. Scarce Societal Resource Allocation and the Price of (Local) Justice. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5628–5636, Virtual Conference, 2021. AAAI Press.

[138] Ziad Obermeyer and Sendhil Mullainathan. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. 2019.

[139] U.S. Dept. of Housing, Office of Policy Development Urban Development, and Research. The applicability of housing first models to homeless persons with serious mental illness: Final report. Technical report, OFFICE OF POLICY DEVELOPMENT AND RESEARCH (PD&R), 2007.

[140] Orgcode. Transition Age Youth – Vulnerability Index – Service Prioritization Decision Assistance Tool (TAY-VI-SPDAT): Next Step Tool for Homeless Youth. Technical report, http://ctagroup.org/wp-content/ uploads/2015/10/Y-SPDAT-v1.0-Youth-Print.pdf, 2015.

[141] OrgCode. The Time Seems Right: Let's Begin the End of the VI-SPDAT, 12 2020.

[142] James B. Orlin, Andreas Schulz, and Rajan Udwani. Robust monotone submodular function maximization. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 312–324, Waterloo, Canada, 2016. Springer.

[143] Derek Parfit. Equality and priority. *Ratio*, 10(3):202–221, 1997.

[144] Douglas Paton and David Johnston. *Disaster resilience: an integrated approach.* Charles C Thomas Publisher, 2017.

[145] Arthur Pigou. *Wealth and welfare.* Macmillan and Company, 1912.

[146] Krzysztof Postek and Dick den Hertog. Multistage adjustable robust mixed-integer optimization via iterative splitting of the uncertainty set. *INFORMS Journal on Computing*, 28(3):553–574, 2016.

[147] Lincoln Quillian. Measuring Racial Discrimination. *Contemporary Sociology: A Journal of Reviews*, 35(1), 2006.

[148] Aida Rahmattalabi, Shahin Jabbari, Himabindu Lakkaraju, Phebe Vayanos, Max Izenberg, Ryan Brown, Eric Rice, and Milind Tambe. Fair influence maximization: a welfare optimization approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11630–11638, 2021.

[149] Aida Rahmattalabi, Phebe Vayanos, Kathryn Dullerud, and Eric Rice. Learning resource allocation policies from observational data with an application to homeless services delivery. *arXiv preprint arXiv:2201.10053*, 2022.

[150] Aida Rahmattalabi, Phebe Vayanos, Kathryn Dullerud, and Eric Rice. Learning resource allocation policies from observational data with an application to homeless services delivery. *arXiv preprint arXiv:2201.10053*, 2022.

[151] Aida Rahmattalabi, Phebe Vayanos, Anthony Fulginiti, Eric Rice, Bryan Wilder, Amulya Yadav, and Milind Tambe. Exploring algorithmic fairness in robust graph covering problems. In *Advances in Neural Information Processing Systems 32*, pages 15750–15761, 2019.

[152] Aida Rahmattalabi, Phebe Vayanos, Anthony Fulginiti, and Milind Tambe. Robust peer-monitoring on graphs with an application to suicide prevention in social networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2168–2170, 2019.

[153] Aida Rahmattalabi, Phebe Vayanos, and Milind Tambe. A robust optimization approach to designing near-optimal strategies for constant-sum monitoring games. In *International Conference on Decision and Game Theory for Security*, pages 603–622. Springer, 2018.

[154] Aida Rahmattalabi, Phebe Vayanos, and Milind Tambe. A robust optimization approach to designing near-optimal strategies for constant-sum monitoring games. In *International Conference on Decision and Game Theory for Security*, pages 603–622. Springer, 2018.

[155] Ab Rashid Ahmad, Zainal Arsad Md Amin, Che Hassandi Abdullah, and Siti Zarina Ngajam. Public awareness and education programme for landslide management and evaluation using a social research approach to determining "acceptable risk" and "tolerable risk" in landslide risk areas in Malaysia. In Kyoji Sassa, Matjaž Mikoš, and Yueping Yin, editors, *Advancing Culture of Living with Landslides*, pages 437–447. Springer International Publishing, 2017.

[156] John Rawls. *A theory of justice*. Harvard university press, 2009.

[157] Eric Rice. Assessment Tools for Prioritizing Housing Resources for Homeless Youth, 2017.

[158] Eric Rice, Monique Holguin, Hsun-Ta Hsu, Matthew Morton, Phebe Vayanos, Milind Tambe, and Hau Chan. Linking Homelessness Vulnerability Assessments to Housing Placements and Outcomes for Youth. *CITYSCAPE*, 20(3):69–86, 2018.

[159] Lisa Rice and Deidre Swesnik. Discriminatory effects of credit scoring on communities of color, 2012.

[160] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 4 1983.

[161] Donald B. Rubin. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91(434), 1996.

[162] Donald B Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[163] Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1), 2007.

[164] Virginia Sapiro. If U.S. Senator Baker Were A Woman: An Experimental Study of Candidate Images. *Political Psychology*, 3(1/2), 1981.

[165] Erel Segal-Halevi and Warut Suksompong. Democratic fair allocation of indivisible goods. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 482–488. AAAI Press, 2018.

[166] Erel Segal-Halevi and Warut Suksompong. Democratic fair allocation of indivisible goods. *Artificial Intelligence*, 277:103167, 2019.

[167] Amartya Sen. *On economic inequality*. Oxford university press, 1997.

[168] Tom Simonite. Meet the secret algorithm that's keeping students out of college, 2020.

[169] Arvind Narayanan Solon Barocas, Moritz Hardt. Fairness in Machine Learning Limitations and Opportunities. *Book*, 2020.

[170] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1), 2016.

[171] Ana-Andreea Stoica, Jessy Xinyi Han, and Augustin Chaintreau. Seeding network influence in biased networks and the benefits of diversity. In *Proceedings of The Web Conference*, pages 2089–2098, 2020.

[172] Xuanming Su and Stefanos Zenios. Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing & Service Operations Management*, 6(4):280–301, 2004.

[173] Warut Suksompong. Approximate maximin shares for groups of agents. *Mathematical Social Sciences*, 92:40–47, 2018.

[174] William Thomson. Problems of fair division and the egalitarian solution. *Journal of Economic Theory*, 31(2):211–226, 1983.

[175] Alan Tsang, Bryan Wilder, Eric Rice, Milind Tambe, and Yair Zick. Group-fairness in influence maximization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5997–6005, 2019.

[176] Vasileios Tzoumas, Konstantinos Gatsis, Ali Jadbabaie, and George J Pappas. Resilient monotone submodular function maximization. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1362–1367. IEEE, 2017.

[177] United States Interagency Council on Homelessness. Opening doors: Federal strategic plan to prevent and end homelessness. Technical report, US Interagency Council on Homelessness, 2015.

[178] Thomas Valente, Anamara Ritt-Olson, Alan Stacy, Jennifer Unger, Janet Okamoto, and Steve Sussman. Peer acceleration: effects of a social network tailored substance abuse prevention program among high-risk adolescents. *Addiction*, 102(11):1804–1815, 2007.

[179] Tyler J. VanderWeele and Whitney R. Robinson. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, 25(4), 2014.

[180] Hal R Varian. Equity, envy, and efficiency. 1973.

[181] Phebe Vayanos, Angelos Georghiou, and Han Yu. Robust optimization with decision-dependent information discovery. *arXiv preprint arXiv:2004.08490*, 2020.

[182] Phebe Vayanos, Daniel Kuhn, and Bercc Rustem. Decision rules for information discovery in multi-stage stochastic programming. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 7368–7373. IEEE, 2011.

[183] Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[184] Jean Walrand. Lecture notes on probability theory and random processes. 2004.

[185] Bryan Wilder, Laura Onasch-Vera, Graham Diguiseppi, Robin Petering, Chyna Hill, Amulya Yadav, Eric Rice, and Milind Tambe. Clinical trial of an ai-augmented intervention for hiv prevention in youth experiencing homelessness, 2020.

[186] Alice Xiang. Reconciling legal and technical approaches to algorithmic bias. *Tenn. L. Rev.*, 88:649, 2020.

[187] Alice Xiang and Donald B. Rubin. Assessing the potential impact of a nationwide class-based affirmative action system. *Statistical Science*, 30(3), 2015.

[188] Amulya Yadav, Hau Chan, Albert Jiang, Haifeng Xu, Eric Rice, and Milind Tambe. Using social networks to aid homeless shelters: Dynamic influence maximization under uncertainty. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 2016.

[189] İhsan Yanıkouglu, Bram L. Gorissen, and Dick den Hertog. A survey of adjustable robust optimization. *European Journal of Operational Research*, 277(3):799–813, 2019.

[190] Naohiro Yonemoto, Yoshitaka Kawashima, Kaori Endo, and Mitsuhiko Yamada. Gatekeeper training for suicidal behaviors: A systematic review. *Journal of affective disorders*, 246:506–514, 2019.

[191] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[192] Chongjie Zhang and Julie A Shah. Fairness in multi-agent sequential decision-making. *Advances in Neural Information Processing Systems*, 27, 2014.

[193] Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, volume 2018-December, 2018.

[194] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making the causal explanation formula. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018.

Appendix A

Technical Appendix to Chapter 2

A.1 Experimental Results in Section 1.6

Data and Data Preprocessing. The original datasets used throughout our paper are described in detail in [17]. They present 8 racial groups, with each individual belonging to a single group. To avoid misinterpretation of the results, we collect racial groups with a population $< 10\%$ of the network size N under the "Other" category. The racial composition of the networks after the preprocessing is provided in Table A.1. For instance, network `SPY1` consists of 54% White, 11% Black, 15% Mixed and 20% Others. The empty entry for Hispanic indicates that their population was less than 10%; as a result, they are categorized under "Other".

Network Name	White	Black	Hispanic	Mixed	Other
`SPY1`	54	11	–	15	20
`SPY2`	55	–	11	21	13
`SPY3`	58	–	10	18	14
`MFP1`	16	38	22	16	8
`MFP2`	16	32	22	20	10

Table A.1: Racial composition (%) of the social networks considered after preprocessing

Setting of Parameter W. We now describe in detail the procedure we use to select W in our experiments. As noted in Section 1.3, to achieve maximin fairness, W must take the maximum value for which the problem is feasible (fairness constraints satisfied). Its value thus depends on other parameters, including I, J, and K. In our experiments, we conduct a search to identify

the best value of W for each setting. Specifically, we vary W from 0 to 1, in increments of 0.04; we employ the largest W for which the problem is feasible. By construction, this choice of W guarantees that all of the fairness constraints are satisfied. In Table A.2, we provide the values of W associated with the results in Table 1.2 for $I = N/3$ and $K = 3$ and for each of the values of J.

Network Name	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$
SPY1	0.44	0.40	0.36	0.32	0.32
SPY2	0.56	0.52	0.48	0.44	0.36
SPY3	0.44	0.36	0.32	0.28	0.24
MFP1	0.52	0.48	0.44	0.40	0.32
MFP2	0.56	0.52	0.44	0.40	0.32

Table A.2: Values of W output by our search procedure and used in the experiments associated with Table 1.2.

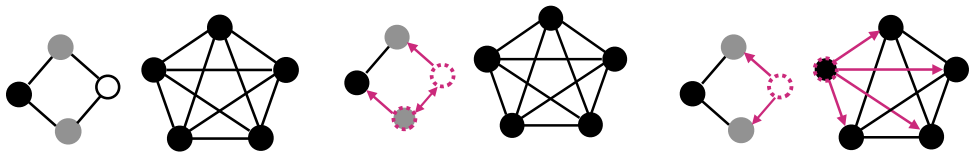
Head-to-Head Comparison with Table 1.1. We conduct a head-to-head comparison of our approach with the results from Table 1.1 which motivated our work. The results are summarized in Table A.3. From the table we observe a consistent increase of 8-14% in worst-case coverage of the worse-off group. For example, in SPY3, the coverage of Hispanics has increased from 33% to 44%. We can also see that the PoF is moderate, ranging from 1-4.2%. The result for the MFP1 network suggests a 36% increase in the coverage of the "Other" group. We note that, by construction, this group consists of racial minorities with a population less than 10% of the network size. While this increase has impacted the coverage of "majority" groups, the worst-case coverage of the worse-off group has increased by 14% with a negligible PoF of 2.6%.

A.2 Proof of Statements in Section 1.3

Proof of Lemma 1. For the special case when all monitors are available ($\Xi = \{\mathbf{e}\}$), there is a single community ($C = 1$), and no fairness constraints are imposed ($W = 0$), Problem ($\mathcal{RC}_{\text{fair}}$) reduces to the maximum coverage problem, which is known to be $\mathcal{NP}$-hard [68]. ■

Network Name	Network Size (N)	Worst-case coverage of individuals by racial group (%)					PoF (%)
		White	Black	Hispanic	Mixed	Other	
SPY1	95	65 (70)	**45** (36)	–	79 (86)	88 (94)	3.3
SPY2	117	81 (78)	–	**50** (42)	72 (76)	73 (67)	1.0
SPY3	118	90 (88)	–	**44** (33)	85 (95)	87 (69)	4.2
MFP1	165	85 (96)	69 (77)	42 (69)	73 (73)	**64** (28)	2.6
MFP2	182	**56** (44)	80 (85)	70 (70)	71 (77)	72 (72)	3.4

Table A.3: Reduction in racial discrimination in node coverage resulting from applying our proposed algorithm relative to that of [176] on the five real-world social networks from Table A.1, when 1/3 of nodes (individuals) can be selected as monitors, out of which at most 10% may fail. The numbers correspond to the worst-case percentage of covered nodes across all monitor availability scenarios. The numbers in the parentheses are solutions to the state-of-the-art algorithm [176] (same numbers as in Table 1.1.



a. Original Graph b. With fairness c. Without fairness

Table A.4: Companion figure to Lemma 2. The figures illustrate a network sequence $\{\mathcal{G}_N\}_{N=5}^{\infty}$ parameterized by N and consisting of two disconnected clusters: a small and a large one, with 4 and $N-4$ nodes, respectively. The small cluster remains intact as N grows. The nodes in the large cluster form a clique. In the figures, each color (white, grey, black) represents a different group and we investigate the price of imposing fairness across these groups. The subfigures show the original graph (a) and an optimal solution when $I = 2$ monitors can be selected in the cases (b) when fairness constraints are not imposed and (c) when fairness constraints are imposed, respectively. It holds that $\text{OPT}^{\text{fair}}(\mathcal{G}_N, 2, 0) = 4$ and $\text{OPT}(\mathcal{G}_N, 2, 0) = N-3$ so that the PoF in $\mathcal{G}_N$ converges to one as N tends to infinity.

A.3 Proofs of Statements in Section 1.4

In all of our analysis, we assume the graphs are undirected. This can be done without loss of generality and the results hold for directed graphs.

A.3.1 Worst-Case PoF

Proof of Lemma 2. Let $\{\mathcal{G}_N\}_{N=5}^{\infty}$ denote the graph sequence shown in Figure A.4(a) (wherein all edges are bidirectional). The network consists of three groups (e.g., racial groups) for which fair treatment is important. Network $\mathcal{G}_N$ consists of two disjoint clusters: one involving four nodes

and a bigger clique containing the remaining $(N-4)$ nodes. Suppose that we can choose $I=2$ nodes as monitors and that all of them are available ($J=0$). Observe that Problem ($\mathcal{RC}_{\text{fair}}$) is feasible only if $0 \leq W \leq (N-3)^{-1}$. For $W=(N-3)^{-1}$, the optimal solution places both nodes in the smaller cluster, see Figure A.4(b). This way, at least one node from each group is covered. The total coverage for the fair solution is then equal to $\text{OPT}^{\text{fair}}(\mathcal{G}_N, 2, 0) = 4$. The maximum achievable coverage under no fairness constraints, however, is obtained by placing one monitor in each cluster, see Figure A.4(c). Thus, the total coverage is equal to $\text{OPT}(\mathcal{G}_N, 2, 0) = N-3$. As a result, $\text{PoF}(\mathcal{G}_N, 2, 0) = 1-4(N-3)^{-1}$ and for $N \geq 4/\epsilon + 3$, it holds that $\text{PoF}(\mathcal{G}_N, 2, 0) \geq 1-\epsilon$. The proof is complete. ■

A.3.2 Supporting Results for the PoF Derivation

In this section, we provide the preliminary results needed in the derivation of the PoF for both the deterministic and robust graph covering problems. First, we provide two results (Lemmas 4 and 5) from the literature which characterize the maximum degree, as well as the expected number of maximum-degree nodes in sparse Erdős Rény graphs [65, 80]. We note that in SBM graphs which are used in our PoF analysis, each community $c \in \mathcal{C}$, when viewed in isolation, is an instance of the Erdős Rényi graph, in which each edge exists independently with probability p_c^{in}. These results are useful to evaluate the coverage of each community $c \in \mathcal{C}$ under the sparsity Assumption 1. Specifically, they enable us to show in Lemma 6 that, in sparse Erdős Rényi graphs, the coverage can be evaluated approximately as the sum of the degrees of the monitoring nodes. Thus, the maximum coverage within each community in an SBM network can obtained by selecting the maximum degree nodes. Lastly, we prove Lemma 8 which will be useful to show that coverage from monitoring nodes in other communities in SBM networks is negligible.

In what follows, we use $\mathbb{G}_{N,p}$ to denote a random instance of Erdős Rény graphs on vertex set $\mathcal{N}(=\{1,\ldots,N\})$, where each edge occurs independently with probability p. Following the notational conventions in [75], we will say that a sequence of events $\{\mathbb{A}_n\}_{n=1}^{N}$ occurs with high

probability if $\lim_{n\to\infty} \mathbb{P}(\mathbb{A}_n) = 1$ and, given a graph $\mathcal{G}$, we let $\Delta(\mathcal{G})$, the maximum degree of vertices of $\mathcal{G}$.

Theorem 4 ([75, Theorem 3.4])**.** *Let $\{\mathbb{G}_{N,p}\}_{N=1}^{\infty}$ a sequence of graphs. If $p = \Theta(N^{-1})$, then with high probability*

$$\lim_{N\to\infty} \Delta(\mathbb{G}_{N,p}) = \frac{\log N}{\log\log N}.$$

Lemma 5. *Let $\{\mathbb{G}_{N,p}\}_{N=1}^{\infty}$ a sequence of graphs with $p = \Theta(N^{-1})$. Let $\sigma(N) := \log N(\log\log N)^{-1}$. Then, it holds that*

$$\mathbb{E}[X_{\sigma(N)}(\mathbb{G}_{N,p})] \geq N^{\frac{\log\log\log N - o(1)}{\log\log N}},$$

where $X_{\sigma(N)}(\mathbb{G}_{N,p})$ is the number of vertices of degree $\sigma(N)$ in $\mathbb{G}_{N,p}$.

Proof. We borrow results from [75, Theorem 3.4], where the authors show that

$$\mathbb{E}[X_{\sigma(N)}(\mathbb{G}_{N,p})] = \exp\left(\frac{\log N}{\log\log N}\left(\log\log\log N - o(1)\right) + O\left(\frac{\log N}{\log\log N + 2\log\log\log N}\right)\right),$$

We further simplify the expression in Lemma 5 by eliminating the $O(.)$ term and we obtain

$$\mathbb{E}[X_{\sigma(N)}(\mathbb{G}_{N,p})] \geq N^{\frac{\log\log\log N - o(1)}{\log\log N}},$$

■

Lemma 5 ensures that our budget for selecting monitors $I = O(\log N)$, is (asymptotically) smaller than number of nodes with degree $\Delta(\mathbb{G}_{N,p})$.

Lemma 6. *Let $\{\mathbb{G}_{N,p}\}_{N=1}^{\infty}$ be a sequence of graphs with $p = \Theta(N^{-1})$. Suppose that the number of monitors is $I = O(\log N)$. Then, for all ν, there exists a graph $\mathbb{G}_{N,p}$ such that the difference*

between the expected maximum coverage in $\mathbb{G}_{N,p}$ and the expected number of neighbors of the monitoring nodes is bounded. Precisely, if $\boldsymbol{x}(\mathbb{G}_{N,p})$ is the indicator vector of the highest degree nodes in $\mathbb{G}_{N,p}$, we have

$$\sum_{n\in\mathcal{N}} \mathbb{E}\left[\boldsymbol{x}_n(\mathbb{G}_{N,p})|\delta_{\mathbb{G}_{N,p}}(n)|\right] - \mathbb{E}\left[F_{\mathbb{G}_{N,p}}(\boldsymbol{x}(\mathbb{G}_{N,p}),\mathbf{e})\right] \;\leq\; \nu,$$

where $\delta_{\mathbb{G}_{N,p}}(n)$ is the set of neighbors of n in $\mathbb{G}_{N,p}$ and ν is the error term and it is $\nu = o(1)$.

Proof. Let Y_n be the event that node n is covered. Also, let Z_n^i the event that node n is covered by the ith highest degree node (and by potentially other nodes too). Without loss of generality, assume that the nodes with lower indexes have higher degrees, i.e., $|\delta(1)| \geq \cdots \geq |\delta(N)|$. The probability that node n is covered can be written as

$$\mathbb{P}(Y_n) = \mathbb{P}\left(\cup_{i=1}^{I} Z_n^i\right). \tag{A.1}$$

From the Bonferroni inequalities, we have

$$\mathbb{P}(\cup_{i=1}^{I} Z_n^i) \;\geq\; \sum_{i=1}^{I}\left(\mathbb{P}(Z_n^i) - \sum_{j=i}^{I}\mathbb{P}(Z_n^i \cap Z_n^j)\right) \tag{A.2}$$

and

$$\mathbb{P}(\cup_{i=1}^{I} Z_n^i) \;\leq\; \sum_{i=1}^{I}\mathbb{P}(Z_n^i). \tag{A.3}$$

Define $Y := \sum_{i=1}^{N} Y_n$ as the (random) total coverage. With a slight abuse of notation, we view Y_n and Z_n^i as Bernoulli random binary variables that are equal to 1 if and only if the associated event occurs. As a result, we can substitute the probability terms with their expected values. Combining Equations (A.1), (A.2) and (A.3), we obtain

$$\sum_{i=1}^{I}\left(\mathbb{E}[Z_n^i] - \sum_{j=i}^{I}\mathbb{E}[Z_n^i Z_n^j]\right) \;\leq\; \mathbb{E}[Y_n] \;\leq\; \sum_{i=1}^{I}\mathbb{E}[Z_n^i], \quad \forall n \in \mathcal{N},$$

where we used the fact that $\mathbb{P}(Z_n^i \cap Z_n^j) = \mathbb{P}(Z_n^i)\mathbb{P}(Z_n^j) = \mathbb{E}(Z_n^i)\mathbb{E}(Z_n^j) = \mathbb{E}(Z_n^i Z_n^j)$ by independence of the events Z_n^i and Z_n^j. Summing over all n yields

$$\sum_{n\in\mathcal{N}} \left(\sum_{i=1}^{I} \mathbb{E}[Z_n^i] - \sum_{j=i}^{I} \mathbb{E}[Z_n^i Z_n^j] \right) \;\leq\; \sum_{n\in\mathcal{N}} \mathbb{E}[Y_n] \;\leq\; \sum_{n\in\mathcal{N}} \sum_{i=1}^{I} \mathbb{E}[Z_n^i].$$

Changing the order of the summations, it follows that

$$\sum_{i=1}^{I} \left(\sum_{n\in\mathcal{N}} \mathbb{E}[Z_n^i] - \sum_{j=i}^{I} \sum_{n\in\mathcal{N}} \mathbb{E}[Z_n^i Z_n^j] \right) \;\leq\; \mathbb{E}[Y] \;\leq\; \sum_{i=1}^{I} \sum_{n\in\mathcal{N}} \mathbb{E}[Z_n^i],$$

where we have used $\mathbb{E}[Y] = \sum_{i=1}^{N} \mathbb{E}[Y_n]$. By definition of $\delta_{\mathbb{G}_{N,p}}(i)$, since $x_i(\mathbb{G}_{N,p}) = 1$ for $i = 1, \ldots, I$, it holds that the number of nodes covered by node i, $\sum_{n\in\mathcal{N}} \mathbb{E}[Z_n^i] = \mathbb{E}[|\delta_{\mathbb{G}_{N,p}}(i)|]$. Also, we remark that $\mathbb{E}[Y] = \mathbb{E}[F_{\mathbb{G}_{N,p}}(\boldsymbol{x}(\mathbb{G}_{N,p}), \mathrm{e})]$. Thus, the above sequence of inequalities is equivalent to

$$\sum_{i=1}^{I} \left(\mathbb{E}[|\delta_{\mathbb{G}_{N,p}}(i)|] - \sum_{j=i}^{I} \sum_{n\in\mathcal{N}} \mathbb{E}[Z_n^i Z_n^j] \right) \;\leq\; \mathbb{E}[F_{\mathbb{G}_{N,p}}(\boldsymbol{x}(\mathbb{G}_{N,p}), \mathbf{e})] \;\leq\; \sum_{i=1}^{I} \mathbb{E}[|\delta_{\mathbb{G}_{N,p}}(i)|],$$

where, by reordering terms, we obtain

$$0 \;\leq\; \sum_{i=1}^{I} \mathbb{E}[|\delta_{\mathbb{G}_{N,p}}(i)|] - \mathbb{E}[F_{\mathbb{G}_{N,p}}(\boldsymbol{x}(\mathbb{G}_{N,p}), \mathbf{e})] \;\leq\; \sum_{i=1}^{I} \sum_{j=i}^{I} \sum_{n\in\mathcal{N}} \mathbb{E}[Z_n^i Z_n^j].$$

Note that $\mathbb{E}\left[\boldsymbol{x}_n(\mathbb{G}_{N,p})\right] = 1, \forall n \leq I$ since by assumption the nodes are ordered by decreasing order of their degree, so the nodes indexed from 1 to I are selected in each realization of the graph. Thus,

$$\begin{aligned} \sum_{i=1}^{I} \mathbb{E}[|\delta_{\mathbb{G}_{N,p}}(i)|] &= \sum_{n\in\mathcal{N}} \mathbb{E}\left[\boldsymbol{x}_n(\mathbb{G}_{N,p})\right] \mathbb{E}\left[|\delta_{\mathbb{G}_{N,p}}(n)|\right] \\ &= \sum_{n\in\mathcal{N}} \mathbb{E}\left[\boldsymbol{x}_n(\mathbb{G}_{N,p}) |\delta_{\mathbb{G}_{N,p}}(n)|\right], \end{aligned}$$

which yields

$$\sum_{n\in\mathcal{N}} \mathbb{E}\left[\boldsymbol{x}_n(\mathbb{G}_{N,p})|\delta_{\mathbb{G}_{N,p}}(n)|\right] - \mathbb{E}[F_{\mathbb{G}_{N,p}}(\boldsymbol{x}(\mathbb{G}_{N,p}),\mathbf{e})] \;\leq\; \sum_{i=1}^{I}\sum_{j=i}^{I}\sum_{n\in\mathcal{N}} \mathbb{E}[Z_n^i Z_n^j]. \tag{A.4}$$

The right-hand side of Equation (A.4) is the error term. We denote this term by ν. This error term determines the difference between the true value of the coverage and the expected sum of the degrees of the monitoring nodes. Given that $p = \Theta(N^{-1})$, we can precisely evaluate the error term. First, we note that since in the Erdős-Rényi model edges are drawn independently, we can write $\mathbb{E}[Z_n^i Z_n^j] = \mathbb{E}[Z_n^i]\,\mathbb{E}[Z_n^j]$. Using Theorem 4 and Lemma 5, and given that the monitors are the highest degree nodes in any realization of the graph, we can write

$$\mathbb{E}[Z_n^i] = \mathbb{E}[Z_n^j] = \Theta\left(\frac{1}{N}\frac{\log N}{\log\log N}\right).$$

We thus obtain

$$\nu = \Theta\left(\frac{I^2}{N}\left(\frac{\log N}{\log\log N}\right)^2\right).$$

By the assumption on the order of I, it follows that $\lim_{N\to\infty}\nu = 0$, which concludes the proof. ■

We now prove the following lemma which will be used in proof of the subsequent results.

Lemma 7. *Let X_i for $i = 1,\ldots,Q$ be Q i.i.d samples from normal distribution with mean μ and standard deviation σ. Also, let $Z = \max_{i\in\{1,\cdots,Q\}} X_i$. It holds that*

$$\mathbb{E}[Z] \leq \mu + \sigma\sqrt{2\log Q}.$$

Proof. By Jensen's inequality,

$$\begin{aligned}\exp(t\,\mathbb{E}[Z]) \leq \mathbb{E}[\exp(tZ)] &= \mathbb{E}[\exp(t \max_{i=1,\ldots,Q} X_i)] \\ &\leq \sum_{i=1}^{Q} \mathbb{E}[\exp(tX_i)] \\ &= Q\exp(\mu t + t^2\sigma^2/2),\end{aligned}$$

where the last equality follows from the definition of the Gaussian moment generating function. Taking the logarithm of both sides of this inequality, we can obtain

$$\mathbb{E}[Z] \leq \mu + \frac{\log Q}{t} + \frac{t\sigma^2}{2}.$$

For the tightest upper-bound, we set $t = \sqrt{2\log Q}/\sigma$. Thus, we obtain

$$\mathbb{E}[Z] \leq \mu + \sigma\sqrt{2\log Q}.$$

■

Lemma 8. *Consider $\mathbb{B}_{N,M,p}$ to be a random instance of a bipartite graph on the vertex set $\mathcal{N} = \mathcal{L} \cup \mathcal{R}$, where $N = |\mathcal{R} \cup \mathcal{L}|$ and $M := |\mathcal{R}|$ and $p = O\left((M\log^2 M)^{-1}\right)$ is the probability that each edge exists (independently). Suppose that monitoring nodes can only be chosen from the set $\mathcal{L}$ and that at most I monitors can be selected. Then, it holds that*

$$\mathbb{E}\left[\max_{\substack{\boldsymbol{x}\in\{0,1\}^{|\mathcal{L}|}:\\ \sum_{n\in\mathcal{L}}\boldsymbol{x}_n = I}} F_{\mathbb{B}_{N,M,p}}(\boldsymbol{x},\mathbf{e})\right] = IO\left(\frac{1}{\log^2 M}\right).$$

Proof. We note that the degree of node i, $\delta_{\mathbb{B}_{N,M,p}}(i)$, follows a binomial distribution with mean Mp. Given we are interested in $N, M \to \infty$, we can approximate the binomial distribution with

a normal distribution [184] with mean Mp and standard deviation $\sqrt{Mp(1-p)}$. Using the result of Lemma 7, we obtain

$$\mathbb{E}[\Delta_{\mathbb{B}_{N,M,p}}] = O\left(Mp + \sqrt{Mp(1-p)}\sqrt{2\log\left(N-M\right)}\right) = O(Mp).$$

Using the above result combined with the assumption on p, we can bound the expected maximum degree of $\mathcal{B}$.

$$\mathbb{E}[\Delta_{\mathbb{B}_{N,M,p}}] = O\left(\frac{1}{\log^2 M}\right).$$

As a result, the maximum expected coverage of the I monitoring nodes is upper-bounded as

$$\mathbb{E}\left[\max_{\substack{\boldsymbol{x}\in\{0,1\}^N:\\ \sum_{n\in\mathcal{L}} x_n = I}} F_{\mathbb{B}_{N,M,p}}(\boldsymbol{x}, \mathrm{e})\right] \;\leq\; I\,\mathbb{E}[\Delta_{\mathbb{B}_{N,M,p}}] \;=\; IO\left(\frac{1}{\log^2 M}\right).$$

and the proof is complete. ■

A.3.3 PoF in the Deterministic Case

Next, we prove the main result which is the derivation of the PoF for the deterministic graph covering problem. The idea of the proof is as follows: by Lemmas 5 and 6, we are able to evaluate the coverage of each community. By Lemma 8, we upper bound the between-community coverage. In other words, based on Lemma 8, we conclude that in every instance of the coverage problem, the between-community coverage is zero (asymptotically) with high probability. Thus, the allocation of monitoring nodes is only dependant on the within-community coverage. Using this observation, we can determine the allocation of the monitors both in the presence and absence of fairness constraints. Subsequently, we are able to evaluate the coverage in both cases. PoF can be then computed based on these two quantities, see Equation (1.2).

Proof of Proposition 1. Let $\mathbb{S}_N$ be a random instance of the SBM network with size N. Consider $\boldsymbol{s}(\mathbb{S}_N) \in \mathbb{Z}^C$ to be the number of allocated monitoring nodes to each of the C communities, i.e., $\boldsymbol{s}_c(\mathbb{S}_N) = \sum_{n \in \mathcal{N}_c} \boldsymbol{x}_n(\mathbb{S}_N)$. Using the result of Lemmas 6 and 8, we can measure the expected maximum coverage as

$$\lim_{N\to\infty} \mathbb{E}[\mathrm{OPT}(\mathbb{S}_N, I, 0)] = \lim_{N\to\infty} \mathbb{E}\left[\max_{\boldsymbol{x}(\mathbb{S}_N)\in\mathcal{X}} F_{\mathbb{S}_N}(\boldsymbol{x}, \mathbf{e})\right] = \mathbb{E}\left[\lim_{N\to\infty} \max_{\boldsymbol{x}(\mathbb{S}_N)\in\mathcal{X}} F_{\mathbb{S}_N}(\boldsymbol{x}, \mathbf{e})\right],$$

where the last equality is obtained by exchanging the expectation and limit. Using Lemma 4 and since the maximum degree is convergent to $d(c)$, we can exchange the limit and maximization term. Thus, we will have

$$\begin{aligned}
\mathbb{E}\left[\lim_{N\to\infty} \max_{\boldsymbol{x}(\mathbb{S}_N)\in\mathcal{X}} F_{\mathbb{S}_N}(\boldsymbol{x}, \mathbf{e})\right] &= \mathbb{E}\left[\max_{\boldsymbol{x}(\mathbb{S}_N)\in\mathcal{X}} \lim_{N\to\infty} F_{\mathbb{S}_N}(\boldsymbol{x}, \mathbf{e})\right] \\
&= \mathbb{E}\left[\max_{\boldsymbol{s}(\mathbb{S}_N)\in\mathbb{Z}^C} \sum_{c\in\mathcal{C}} \boldsymbol{s}_c(\mathbb{S}_N) d(c) + o(1)\right],
\end{aligned}$$

which given that $d(c)$ is only dependent on the size of the communities in $\mathbb{S}_N$ is equivalent to

$$\lim_{N\to\infty} \mathbb{E}[\mathrm{OPT}(\mathbb{S}_N, I, 0)] = \max_{\boldsymbol{s}(\mathbb{S}_N)} \sum_{c\in\mathcal{C}} \boldsymbol{s}_c(\mathbb{S}_N) d(c) + o(1). \tag{A.5}$$

Equation (A.5) suggests that for large enough N, the maximum coverage is only dependent on the *number* of the monitoring nodes allocated to each community. Also, the allocation is the same for all random instances so we can drop the dependence of $\boldsymbol{s}$ on $\mathbb{S}_N$. In right-hand side of Equation (A.5), the first term is the within-community (Lemma 6), and the second term is the between-community (Lemma 8) coverage.

In the analysis below, all the evaluations are for large enough N. Therefore, we drop the $\lim_{N\to\infty}$ for ease of notation. According to Equation (A.5) the between-community coverage is negligible, compared to the within-community coverage. This suggests that the maximum

achievable coverage will be obtained by placing all the monitoring nodes in the largest community, with the largest value of $d(c)$, where the assumption on I, as given in the premise of the proposition, combined with Lemma 5 guarantee that such a selection is possible. Thus, we obtain

$$\mathbb{E}[\mathrm{OPT}(\mathbb{S}_N, I, 0)] = Id(C) + o(1).$$

Next, we measure $\mathbb{E}[\mathrm{OPT}^{\mathrm{fair}}(.)]$, where in addition to optimization problem in Equation (A.5), the allocation is further restricted to satisfy all the fairness constraints.

$$\frac{s_c}{|\mathcal{N}_c|} d(c) + o(1) \geq W \quad \forall c \in \mathcal{C}, \tag{A.6}$$

in which, $o(1)$ is the term that compensates for the coverage of the nodes in other communities, and is small due to the regimes of $p_{cc'}^{\mathrm{out}},\ \forall c, c' \in \mathcal{C}$ and the budget I. At optimality and for the maximum value of W, we have

$$\left| s_c |\mathcal{N}_c|^{-1} d(c) - s_{c'} |\mathcal{N}_{c'}|^{-1} d(c') \right| \leq \delta \ \ \forall c, c' \in \mathcal{C}, \delta \leq \left| d(1)|\mathcal{N}_1|^{-1} - d(C)|\mathcal{N}_C|^{-1} \right|.$$

This holds because otherwise one can remove on node from the group with higher value of $s_c|\mathcal{N}_c|^{-1}d(c)$ to a group with less value and thus increase the normalized coverage of the worse-off group and this contradicts the fact that W is the maximum possible value. This suggests that in a fair solution, the normalized coverage is *almost* equal across different groups, given that $\lim_{N\to\infty} \delta = 0$. As a result, the monitoring nodes should be such that

$$W \leq \frac{s_c}{|\mathcal{N}_c|} d(c) + o(1) \leq W + \delta,\ \forall c \in \mathcal{C}.$$

From this, it follows that

$$W - o(1) \leq \frac{s_c}{|\mathcal{N}_c|} d(c) \leq W + o(1). \tag{A.7}$$

By assumption, there must be an integral s_c that satisfies the above relation. Note that if we could relax the integrality assumption, $s_c = W|\mathcal{N}_c|d(c)^{-1}$. Due to the integrality constraint, and according to Equation (A.7), we set $s_c|\mathcal{N}_c|^{-1}d(c) = W + o(1)$, where the o(1) term is to account for the discretizing error, which results in $s_c = W|\mathcal{N}_c|d(c)^{-1} + O(1)$, where $O(1) \leq 1$ (As we can not make a higher error in rounding). Also, since $\sum_{c\in\mathcal{C}} s_c = I$, we can obtain the value of W as

$$W = \frac{I}{\sum_{c\in\mathcal{C}} \frac{|\mathcal{N}_c|}{d(c)}} + o(1).$$

As a result

$$s_c = \frac{I}{\sum_{c\in\mathcal{C}} \frac{|\mathcal{N}_c|}{d(c)}} \frac{|\mathcal{N}_c|}{d(c)} + O(1) \quad \forall c \in \mathcal{C}.$$

We now define $\kappa := I \left(\sum_{c\in\mathcal{C}} \frac{|\mathcal{N}_c|}{d(c)}\right)^{-1}$ for a compact representation.

So far, we obtained the allocation of the monitoring nodes to satisfy the fairness constraints. This is enough to evaluate the coverage under the fairness constraints. Now, we can evaluate the PoF as defined by Equation (1.2).

$$
\begin{aligned}
& & \mathbb{E}[\mathrm{OPT}(\mathbb{S}_N, I, 0)] &= I d(C) \\
\Rightarrow & & -\frac{1}{\mathbb{E}[\mathrm{OPT}(\mathbb{S}_N, I, 0)]} &= -\frac{1}{I\, d(C)} \\
\Rightarrow & & -\frac{\mathbb{E}[\mathrm{OPT}^{\mathrm{fair}}(\mathbb{S}_N, I, 0)]}{\mathbb{E}[\mathrm{OPT}(\mathbb{S}_N, I, 0)]} &= -\frac{\kappa \sum_{c\in\mathcal{C}} \frac{|\mathcal{N}_c|}{d(c)} d(c)}{I\, d(C)} - o(1) \\
\Rightarrow & 1 - \frac{\mathbb{E}[\mathrm{OPT}^{\mathrm{fair}}(\mathbb{S}_N, I, 0)]}{\mathbb{E}[\mathrm{OPT}(\mathbb{S}_N, I, 0)]} &= 1 - \frac{\kappa \sum_{c\in\mathcal{C}} \frac{|\mathcal{N}_c|}{d(c)} d(c)}{I\, d(C)} - o(1) \\
\Rightarrow & & \overline{\mathrm{PoF}}(I, 0) &= 1 - \frac{\kappa \sum_{c\in\mathcal{C}} |\mathcal{N}_c|}{I\, d(C)} - o(1) \\
\Rightarrow & & \overline{\mathrm{PoF}}(I, 0) &= 1 - \frac{\sum_{c\in\mathcal{C}} |\mathcal{N}_c|}{\sum_{c\in\mathcal{C}} |\mathcal{N}_c| d(C)/d(c)} - o(1).
\end{aligned}
$$

■

A.3.4 PoF in the Robust Case

Proof of Proposition 2. The idea of the proof is similar to Proposition 1, with the exception that the fair allocation of the monitoring nodes will be affected by the uncertainty. Consider $\boldsymbol{s}$ to be the number of allocated monitoring nodes to each of the C communities, i.e., $s_c = \sum_{n\in\mathcal{N}_c} x_n$. Using the result of lemma 6, and 8, we can measure the expected maximum coverage as

$$\mathbb{E}[\mathrm{OPT}(\mathbb{S}_N, I, J)] = (I - J)d(c) + o(1).$$

That is because, in the worst-case J nodes fail, thus only $(I - J)$ nodes can cover the graph. Next, we measure $\mathbb{E}[\text{OPT}^{\text{fair}}(.)]$, where in addition to optimization problem in Equation (A.5), the allocation is further restricted to satisfy all the fairness constraints. Given that at most J nodes may fail, we need to ensure after fairness constraints are satisfied after the removal of J nodes. We momentarily revisit the fairness constraint in the deterministic case.

$$\frac{s_c}{|\mathcal{N}_c|} d(c) + o(1) \geq W \quad \forall c \in \mathcal{C},$$

in which, $o(1)$ is the term that compensates for the coverage of the nodes in other communities, and is small due to the regimes of p^{out}, and the budget I. Under the uncertainty, we need to ensure that these constraints are satisfied even after J nodes are removed. In other words

$$\frac{(s_c - J)}{|\mathcal{N}_c|} d(c) + o(1) \geq W \quad \forall c \in \mathcal{C}.$$

At optimality and for the maximum value of W, we have

$$\left|(s_c - J)|\mathcal{N}_c|^{-1} d(c) - (s_{c'} - J)|\mathcal{N}_{c'}|^{-1} d(c')\right| \leq \delta \;\; \forall c, c' \in \mathcal{C}, \delta \leq \left|d(1)|\mathcal{N}_1|^{-1} - d(C)|\mathcal{N}_C|^{-1}\right|.$$

This holds because otherwise one can remove on node from the group with higher value of $s_c|\mathcal{N}_c|^{-1} d(c)$ to a group with less value and thus increase the normalized coverage of the worse-off group and this contradicts the fact that W is the maximum possible value.

This suggests that in a fair solution, the normalized coverage is *almost* equal across different groups, given that $\delta \to 0$, as $\mathcal{N}_c \to \infty, \forall c \in \mathcal{C}$. Following the proof of Proposition 1, the discretizing error can be handled by setting $(s_c - J)|\mathcal{N}_c|^{-1} d(c) = W + o(1)$, where the o(1) term is to account for the discretizing error. As a result

$$s_c = \frac{|\mathcal{N}_c| W}{d(c)} + J + O(1),$$

where $O(1) \leq 1$ (As we can not make a higher error in rounding). This suggests that a fair allocation is the one that places J nodes in each community, regardless of the community size. The remaining monitors are allocated with respect to the relative size of the communities.

Summing over all s_c and since $\sum_{c\in\mathcal{C}} s_c = I$ we obtain

$$W = \frac{(I - CJ)}{\sum_{c\in\mathcal{C}} \frac{|\mathcal{N}_c|}{d(c)}} + o(1).$$

As a result

$$s_c = \frac{(I - CJ)}{\sum_{c\in\mathcal{C}} \frac{|\mathcal{N}_c|}{d(c)}} \frac{|\mathcal{N}_c|}{d(c)} + J + O(1) \quad \forall c \in \mathcal{C}.$$

As defined in the premise of the proposition, $\eta = (I - CJ)\left(\sum_{c\in\mathcal{C}} \frac{|\mathcal{N}_c|}{d(c)}\right)^{-1}$.

So far, we obtained the allocation of the monitoring nodes, to satisfy the fairness constraints.

Now, we evaluate the coverage, i.e., objective value of Problem ($\mathcal{RC}_{\text{fair}}$), under the obtained fair allocation. Since the fairness constraints are satisfied under all the scenarios, the worst-case scenario is the one that results in the maximum loss in the total coverage. This corresponds to the case that J nodes from the largest community ($\mathcal{N}_C$) fail. As a result the expected coverage can be obtained by

$$\mathbb{E}[\text{OPT}^{\text{fair}}(\mathbb{S}_N, I, J)] = \sum_{c\in\mathcal{C}} \left(\eta \frac{|\mathcal{N}_c|}{d(c)} d(c) + Jd(c) + O(1)d(c)\right) - Jd(C).$$

Now, we can evaluate the PoF as defined by Equation (1.2).

$$\begin{aligned}
& \mathbb{E}[\text{OPT}(\mathbb{S}_N, I, J)] &=& \ (I - J)d(C) \\
\Rightarrow \quad & -\frac{1}{\mathbb{E}[\text{OPT}(\mathbb{S}_N, I, J)]} &=& -\frac{1}{(I - J)d(C)} \\
\Rightarrow \quad & -\frac{\mathbb{E}[\text{OPT}^{\text{fair}}(\mathbb{S}_N, I, J)]}{\mathbb{E}[\text{OPT}(\mathbb{S}_N, I, J)]} &=& -\frac{\sum_{c\in\mathcal{C}} (\eta|\mathcal{N}_c| + Jd(c)) - Jd(C)}{(I - J)d(C)} - o(1) \\
\Rightarrow \quad & 1 - \frac{\mathbb{E}[\text{OPT}^{\text{fair}}(\mathbb{S}_N, I, J)]}{\mathbb{E}[\text{OPT}(\mathbb{S}_N, I, J)]} &=& 1 - \frac{\sum_{c\in\mathcal{C}} \eta|\mathcal{N}_c| + \sum_{c\in\mathcal{C}\setminus\{C\}} Jd(c)}{(I - J)d(C)} - o(1) \\
\Rightarrow \quad & \overline{\text{PoF}}(I, J) &=& \ 1 - \frac{\sum_{c\in\mathcal{C}} \eta|\mathcal{N}_c|}{(I - J)d(C)} - \frac{J\sum_{c\in\mathcal{C}\setminus\{C\}} d(c)}{(I - J)d(C)} - o(1).
\end{aligned}$$

■

A.4 Proofs of Statements in Section 1.5

A.4.1 Equivalent Reformulation as a Max-Min-Max Optimization

Proof of Proposition 3. Let $\bar{x}$ be feasible in Problem ($\mathcal{RC}_{\text{fair}}$). It follows that it is also feasible in Problem 1.3. For a fixed $\bar{\xi}$, we show that

$$\begin{aligned}\sum_{c\in\mathcal{C}} F_{\mathcal{G},c}(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}}) \quad = \quad & \max_{\boldsymbol{y}} \quad \sum_{c\in\mathcal{C}}\sum_{n\in\mathcal{N}_c} \boldsymbol{y}_n \\ & \text{s.t.} \quad \boldsymbol{y}_n \leq \sum_{\nu\in\delta(n)} \bar{\boldsymbol{\xi}}_\nu \bar{\boldsymbol{x}}_\nu \\ & \qquad \sum_{n\in\mathcal{C}} \boldsymbol{y}_n \geq W|\mathcal{N}_c|, \ \forall c\in\mathcal{C}\end{aligned}$$

Since $\bar{\boldsymbol{x}}$ is feasible in Problem ($\mathcal{RC}_{\text{fair}}$), it holds that

$$\begin{aligned} F_{\mathcal{G},c}(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}}) \quad &= \quad \sum_{n\in\mathcal{N}_c} \boldsymbol{y}_n(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}}) \\ &= \quad \sum_{n\in\mathcal{N}_c} \mathbb{I}\left(\sum_{\nu\in\delta(n)} \bar{\boldsymbol{\xi}}_\nu \bar{\boldsymbol{x}}_\nu \geq 1\right) \\ &\geq \quad W|\mathcal{N}_c| \end{aligned}$$

We define $\boldsymbol{y}_n^\star = \mathbb{I}\left(\sum_{\nu\in\delta(n)} \bar{\boldsymbol{\xi}}_\nu \bar{\boldsymbol{x}}_\nu \geq 1\right)$ which is feasible in Problem (1.3). Since the choice of $\bar{\boldsymbol{\xi}}$ was arbitrary, we showed that given a solution to Problem ($\mathcal{RC}_{\text{fair}}$), we can always construct a feasible solution to Problem (1.3), thus the objective value of the latter is at least as high.

We now prove the contrary, i.e., given a solution to Problem (1.3), we will construct a solution to Problem ($\mathcal{RC}_{\text{fair}}$). Consider $\bar{\boldsymbol{x}}$ to be an optimal solution to Problem ($\mathcal{RC}_{\text{fair}}$). Suppose there exists $\bar{\boldsymbol{\xi}} \in \Xi$ such that

$$\begin{aligned} F_{\mathcal{G},c}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\xi}}) \quad &< \quad |\mathcal{N}_c|W \\ \Rightarrow \quad \sum_{n \in \mathcal{N}_c} \mathbb{I}\left(\sum_{\nu \in \delta(n)} \bar{\boldsymbol{\xi}}_\nu \bar{\boldsymbol{x}}_\nu \geq 1 \right) \quad &< \quad |\mathcal{N}_c|W. \end{aligned}$$

However, since $\bar{\boldsymbol{x}}$ is feasible in Problem ($\mathcal{RC}_{\text{fair}}$), we have that

$$\begin{aligned} \forall \tilde{\boldsymbol{\xi}} \in \Xi, \; \exists \boldsymbol{y}_n : \quad \boldsymbol{y}_n &\leq \sum_{\nu \in \delta(n)} \tilde{\boldsymbol{\xi}}_\nu \bar{\boldsymbol{x}}_\nu \\ \sum_{n \in \mathcal{N}_c} \boldsymbol{y}_n &\geq |\mathcal{N}_c|W. \end{aligned}$$

By construction, $\boldsymbol{y}_n \leq \mathbb{I}\left(\sum_{\nu \in \delta(n)} \tilde{\boldsymbol{\xi}}_\nu \bar{\boldsymbol{x}}_\nu \geq 1 \right), \; \forall n \in \mathcal{N}$. Thus

$$\begin{aligned} \sum_{c \in \mathcal{C}} \sum_{n \in \mathcal{N}_c} \mathbb{I}\left(\sum_{\nu \in \delta(n)} \tilde{\boldsymbol{\xi}}_\nu \bar{\boldsymbol{x}}_\nu \geq 1 \right) \quad &\geq \quad \sum_{c \in \mathcal{C}} \sum_{n \in \mathcal{N}_c} \boldsymbol{y}_n \\ &\geq \quad |\mathcal{N}_c|W. \end{aligned}$$

According to the above result, we showed that the optimal objective value of Problem ($\mathcal{RC}_{\text{fair}}$) is at least as high as that of Problem (1.3). This completes the proof. ■

A.4.2 Exact MILP Formulation of the K-Adaptability Problem

In order to derive the equivalent MILP in Theorem 1, we start by a variant of the K-adaptability Problem (1.4), in which we move the constraints of the inner maximization problem to the definition of the uncertainty set in the spirit of [84]. Next, we prove, via Proposition 12, that by relaxing the integrality constraint on the uncertain parameters $\boldsymbol{\xi}$, the problem remains unchanged, and this is the key result that enables us to provide an *equivalent* MILP reformulation for Problem (1.4).

We replace Ξ with a collection of uncertainty sets parameterized by vectors $\boldsymbol{\ell} \in \mathcal{L}$ as in [84]. Specifically, it follows from Proposition 2 in [84] that Problem (1.4) is equivalent to

$$\begin{aligned} \max \quad & \min_{\boldsymbol{\ell} \in \mathcal{L}} \quad \min_{\boldsymbol{\xi} \in \Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})} \quad \max_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k = 0}} \quad \sum_{n \in \mathcal{N}} \boldsymbol{y}_n^k \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{y}^1, \ldots, \boldsymbol{y}^K \in \mathcal{Y}, \end{aligned} \tag{A.8}$$

where $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$ is defined through

$$\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) \ := \left\{ \boldsymbol{\xi} \in \Xi : \begin{array}{ll} \boldsymbol{y}_{\boldsymbol{\ell}_k}^k \ > \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu, & \forall k \in \mathcal{K} : \boldsymbol{\ell}_k > 0 \\ \\ \boldsymbol{y}_n^k \ \leq \sum_{\nu \in \delta(n)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu \ \ \forall n \in \mathcal{N}, & \forall k \in \mathcal{K} : \boldsymbol{\ell}_k = 0 \end{array} \right\},$$

and, with a slight abuse of notation, we use $\boldsymbol{y} := \{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^K\}$. The vector $\boldsymbol{\ell} \in \mathcal{L}$ encodes which of the K candidate covering schemes are feasible. By introducing $\boldsymbol{\ell}$, the constraints of the inner maximization problem are absorbed in the parameterized uncertainty sets $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$, and in the inner-most maximization problem, any covering scheme can be chosen for which $\boldsymbol{\ell}_k = 0$.

Note that, for any fixed $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}^K$, and $\boldsymbol{\ell} \in \mathcal{L}$, the strict inequalities in $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$ can be converted to (loose) inequalities as in

$$\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) = \left\{ \boldsymbol{\xi} \in \Xi : \begin{array}{ll} \boldsymbol{y}^k_{\boldsymbol{\ell}_k} \geq \sum\limits_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu + 1, & \forall k \in \mathcal{K} : \boldsymbol{\ell}_k > 0 \\ \boldsymbol{y}^k_n \leq \sum\limits_{\nu \in \delta(n)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu & \forall n \in \mathcal{N},\ \forall k \in \mathcal{K} : \boldsymbol{\ell}_k = 0 \end{array} \right\}.$$

This idea was previously leveraged in [154]. It follows naturally since all decision variables and uncertain parameters are binary. Next, we show that we can obtain an equivalent problem by relaxing the integrality constraint on the set Ξ in the definition of $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{l})$. Consider the following problem

$$\begin{array}{ll} \max & \min\limits_{\boldsymbol{\ell} \in \mathcal{L}} \quad \min\limits_{\boldsymbol{\xi} \in \overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})} \quad \max\limits_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k = 0}} \quad \sum\limits_{n \in \mathcal{N}} \boldsymbol{y}^k_n \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{y} \in \mathcal{Y}^K, \end{array} \tag{A.9}$$

where the uncertainty set is obtained by relaxing the integrality constraints on $\boldsymbol{\xi}$, i.e.,

$$\overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) = \left\{ \boldsymbol{\xi} \in \mathcal{T} : \begin{array}{ll} \boldsymbol{y}^k_{\boldsymbol{\ell}_k} \geq \sum\limits_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu + 1, & \forall k \in \mathcal{K} : \boldsymbol{\ell}_k > 0 \\ \boldsymbol{y}^k_n \leq \sum\limits_{\nu \in \delta(n)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu & \forall n \in \mathcal{N},\ \forall k \in \mathcal{K} : \boldsymbol{\ell}_k = 0 \end{array} \right\}.$$

Proposition 12. *Under Assumption 3, Problems* (A.8) *and* (A.9) *are equivalent.*

Proof. Let $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}^K$, and $\boldsymbol{\ell} \in \mathcal{L}$. It suffices to show that

$$\min_{\boldsymbol{\xi} \in \Xi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\ell})} \max_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k = 0}} \sum_{n \in \mathcal{N}} \boldsymbol{y}_n^k \qquad \text{and} \qquad \min_{\boldsymbol{\xi} \in \overline{\Xi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\ell})} \max_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k = 0}} \sum_{n \in \mathcal{N}} \boldsymbol{y}_n^k$$

are equivalent. Observe that the these problems have the same objective function. Thus, the two problems have the same optimal objective value if and only if they are either both feasible or both infeasible. As a result, it suffices to show that $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$ is empty if and only if $\overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$ is empty. Naturally, if $\overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) = \emptyset$ then $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) = \emptyset$ since $\mathcal{T}$ is the linear programming relaxation of Ξ. Thus, it suffices to show that the converse also holds, i.e., that if $\overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) \neq \emptyset$, then also $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) \neq \emptyset$.

To this end, suppose that $\overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) \neq \emptyset$ and let $\tilde{\boldsymbol{\xi}} \in \overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$. Then, $\tilde{\boldsymbol{\xi}}$ is such that

$$\begin{aligned}
&\tilde{\boldsymbol{\xi}} \in \mathcal{T}, \\
&\boldsymbol{y}^k_{\boldsymbol{\ell}_k} \;\geq \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \tilde{\boldsymbol{\xi}}_\nu \boldsymbol{x}_\nu + 1 \;\; \forall k \in \mathcal{K} : \boldsymbol{\ell}_k > 0, \\
&\boldsymbol{y}^k_n \;\leq \sum_{\nu \in \delta(n)} \tilde{\boldsymbol{\xi}}_\nu \boldsymbol{x}_\nu \;\; \forall n \in \mathcal{N}, \; \forall k \in \mathcal{K} : \boldsymbol{\ell}_k = 0.
\end{aligned} \tag{A.10}$$

Next, define $\hat{\boldsymbol{\xi}}_n := \lceil \tilde{\boldsymbol{\xi}}_n \rceil \;\forall n \in \mathcal{N}$. We show that $\hat{\boldsymbol{\xi}} \in \Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$. First, note that $\hat{\boldsymbol{\xi}} \geq \tilde{\boldsymbol{\xi}}$ and by Assumption 3, it follows that $\hat{\boldsymbol{\xi}} \in \mathcal{T}$. Moreover, by construction, $\hat{\boldsymbol{\xi}} \in \{0, 1\}^N$. Thus, it follows

that $\hat{\boldsymbol{\xi}} \in \Xi$. Next, we show that the constructed solution $\hat{\boldsymbol{\xi}}$ also satisfies the remaining constraints in $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$. Fix $k \in \mathcal{K}$ such that $\boldsymbol{\ell}_k > 0$. Then, from (A.10) it holds that

$$
\begin{aligned}
& \boldsymbol{y}^k_{\boldsymbol{\ell}_k} \geq \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \tilde{\boldsymbol{\xi}}_\nu \boldsymbol{x}_\nu + 1 \\
\Rightarrow \quad & \boldsymbol{y}^k_{\boldsymbol{\ell}_k} = 1 \text{ and } \tilde{\boldsymbol{\xi}}_\nu \boldsymbol{x}_\nu = 0 \quad \forall \nu \in \delta(\boldsymbol{\ell}_k) \\
\Rightarrow \quad & \boldsymbol{y}^k_{\boldsymbol{\ell}_k} = 1 \text{ and } \tilde{\boldsymbol{\xi}}_\nu = 0 \quad \forall \nu \in \delta(\boldsymbol{\ell}_k) : \boldsymbol{x}_\nu = 1 \\
\Rightarrow \quad & \boldsymbol{y}^k_{\boldsymbol{\ell}_k} = 1 \text{ and } \hat{\boldsymbol{\xi}}_\nu = 0 \quad \forall \nu \in \delta(\boldsymbol{\ell}_k) : \boldsymbol{x}_\nu = 1 \\
\Rightarrow \quad & \boldsymbol{y}^k_{\boldsymbol{\ell}_k} \geq \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \hat{\boldsymbol{\xi}}_\nu \boldsymbol{x}_\nu + 1,
\end{aligned}
$$

where the first and second implication follow since $\boldsymbol{y}$ and $\boldsymbol{x}$ are binary, respectively, and the third implication holds by definition of $\hat{\boldsymbol{\xi}}$,

Next, fix $k \in \mathcal{K}$ such that $\boldsymbol{\ell}_k = 0$. Then, (A.10) yields

$$
\begin{aligned}
& \boldsymbol{y}^k_n \leq \sum_{\nu \in \delta(n)} \tilde{\boldsymbol{\xi}}_\nu \boldsymbol{x}_\nu \quad \forall n \in \mathcal{N} \\
\Rightarrow \quad & \boldsymbol{y}^k_n \leq \sum_{\nu \in \delta(n)} \hat{\boldsymbol{\xi}}_\nu \boldsymbol{x}_\nu \quad \forall n \in \mathcal{N},
\end{aligned}
$$

which follows by definition of $\hat{\boldsymbol{\xi}}$. We have thus constructed $\hat{\boldsymbol{\xi}} \in \Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell})$ and therefore conclude that $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) \neq \emptyset$. Since the choice of $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}^K$, and $\boldsymbol{\ell} \in \mathcal{L}$ was arbitrary, the claim follows. ■

Proposition 12 is key to leverage existing literature to reformulate Problem (1.4) as an MILP. The reformulation is based on [84, 154].

Proof of Theorem 1. Note that the objective function of the Problem (A.8) is identical to

$$\min_{\boldsymbol{\ell}\in\mathcal{L}} \quad \min_{\boldsymbol{\xi}\in\overline{\Xi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\ell})} \quad \left[\max_{\boldsymbol{\lambda}\in\Delta_K(\boldsymbol{\ell})} \sum_{k\in\mathcal{K}} \boldsymbol{\lambda}_k \sum_{n\in\mathcal{N}} \boldsymbol{y}_n^k\right],$$

where $\Delta_K(\boldsymbol{\ell}) := \{\boldsymbol{\lambda} \in \mathbb{R}_+^K : \mathbf{e}^\top \boldsymbol{\lambda} = 1,\ \boldsymbol{\lambda}_k = 0\ \ \forall k \in \mathcal{K} : \boldsymbol{\ell}_k \neq 0\}$. We define $\partial\mathcal{L} := \{\boldsymbol{\ell} \in \mathcal{L} : \boldsymbol{\ell} \not> \mathbf{0}\}$, and $\mathcal{L}_+ := \{\boldsymbol{\ell} \in \mathcal{L} : \boldsymbol{\ell} > \mathbf{0}\}$. We remark that $\Delta_K(\boldsymbol{\ell}) = \emptyset$ if and only if $\boldsymbol{\ell} > \mathbf{0}$. If $\overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) = \emptyset$ for all $\boldsymbol{\ell} \in \mathcal{L}_+$, then the problem is equivalent to

$$\min_{\boldsymbol{\ell}\in\partial\mathcal{L}} \quad \min_{\boldsymbol{\xi}\in\overline{\Xi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\ell})} \quad \left[\max_{\boldsymbol{\lambda}\in\Delta_K(\boldsymbol{\ell})} \sum_{k\in\mathcal{K}} \boldsymbol{\lambda}_k \sum_{n\in\mathcal{N}} \boldsymbol{y}_n^k\right].$$

By applying the classical min-max theorem, we obtain

$$\min_{\boldsymbol{\ell}\in\partial\mathcal{L}} \quad \max_{\boldsymbol{\lambda}\in\Delta_K(\boldsymbol{\ell})} \quad \min_{\boldsymbol{\xi}\in\overline{\Xi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\ell})} \sum_{k\in\mathcal{K}} \boldsymbol{\lambda}_k \sum_{n\in\mathcal{N}} \boldsymbol{y}_n^k.$$

This problem is also equivalent to

$$\max_{\boldsymbol{\lambda}(\boldsymbol{\ell})\in\Delta_K(\boldsymbol{\ell})} \quad \min_{\boldsymbol{\ell}\in\partial\mathcal{L}} \quad \min_{\boldsymbol{\xi}\in\overline{\Xi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\ell})} \sum_{k\in\mathcal{K}} \boldsymbol{\lambda}_k(\boldsymbol{\ell}) \sum_{n\in\mathcal{N}} \boldsymbol{y}_n^k.$$

If on the other hand $\overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) \neq \emptyset$ for some $\boldsymbol{\ell} \in \mathcal{L}_+$, the objective of Problem (A.8) evaluates to $-\infty$.

Using the above results, we can write Problem (A.8) in epigraph form as

$$
\begin{aligned}
\max \quad & \tau \\
\text{s.t.} \quad & \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{y} \in \mathcal{Y}^K,\ \tau \in \mathbb{R},\ \boldsymbol{\lambda}(\boldsymbol{\ell}) \in \Delta_K(\boldsymbol{\ell}),\ \boldsymbol{\ell} \in \partial\mathcal{L} \\
& \tau \ \leq \ \sum_{k \in \mathcal{K}} \boldsymbol{\lambda}_k(\boldsymbol{\ell}) \sum_{n \in \mathcal{N}} \boldsymbol{y}_n^k \quad \forall \boldsymbol{\ell} \in \partial\mathcal{L} : \ \Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) \neq \emptyset \\
& \overline{\Xi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) = \emptyset \quad \forall \boldsymbol{\ell} \in \mathcal{L}_+.
\end{aligned}
\tag{A.11}
$$

We begin by reformulating the semi-infinite constraint associated with $\boldsymbol{\ell} \in \partial\mathcal{L}$ in Problem (A.11). To this end, fix $\boldsymbol{\ell} \in \partial\mathcal{L}$ and consider the linear program

$$
\begin{aligned}
\min \quad & 0 \\
\text{s.t.} \quad & 0 \ \leq \ \boldsymbol{\xi}_n \leq 1 \ \ \forall n \in \mathcal{N} \\
& \boldsymbol{A}^\top \boldsymbol{\xi} \ \geq \ \boldsymbol{b} \\
& \boldsymbol{y}^k_{\boldsymbol{\ell}_k} \ \geq \ \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu + 1 \quad \forall k \in \mathcal{K} \ : \ \boldsymbol{\ell}_k > 0 \\
& \boldsymbol{y}^k_n \ \leq \ \sum_{\nu \in \delta(n)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu \quad \forall n \in \mathcal{N},\ \forall k \in \mathcal{K} \ : \ \boldsymbol{\ell}_k = 0,
\end{aligned}
$$

whose dual reads

$$\max \quad -\mathbf{e}^\top \boldsymbol{\theta}(\boldsymbol{\ell}) + \boldsymbol{b}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) - \sum_{\substack{k\in\mathcal{K}\\ \boldsymbol{\ell}_k\neq 0}} \left(\boldsymbol{y}^k_{\boldsymbol{\ell}_k} - 1\right) \boldsymbol{\nu}_k(\boldsymbol{\ell}) + \sum_{\substack{k\in\mathcal{K}\\ \boldsymbol{\ell}_k= 0}} \sum_{n\in\mathcal{N}} \boldsymbol{y}^k_n \boldsymbol{\beta}^k_n(\boldsymbol{\ell})$$

$$\text{s.t.} \quad \boldsymbol{\theta}(\boldsymbol{\ell}) \in \mathbb{R}^N_+,\ \boldsymbol{\alpha}(\boldsymbol{\ell}) \in \mathbb{R}^R_+,\ \boldsymbol{\beta}^k(\boldsymbol{\ell}) \in \mathbb{R}^N_+,\ \forall k \in \mathcal{K},\ \boldsymbol{\nu}(\boldsymbol{\ell}) \in \mathbb{R}^K_+$$

$$\boldsymbol{\theta}_n(\boldsymbol{\ell}) \ \leq\ \boldsymbol{A}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) + \sum_{\substack{k\in\mathcal{K}\\ \boldsymbol{\ell}_k\neq 0}} \sum_{\nu\in\delta(\boldsymbol{\ell}_k)} \boldsymbol{x}_\nu \boldsymbol{\nu}_k(\boldsymbol{\ell}) - \sum_{\substack{k\in\mathcal{K}\\ \boldsymbol{\ell}_k= 0}} \sum_{\nu\in\delta(n)} \boldsymbol{x}_\nu \boldsymbol{\beta}^k_n(\boldsymbol{\ell}) \quad \forall n \in \mathcal{N}.$$

In Problem (A.11) the constraint associated with each $\boldsymbol{\ell} \in \partial\mathcal{L}$ is satisfied if and only if the objective value of the above dual problem is greater than $\tau - \sum_{k\in\mathcal{K}} \boldsymbol{\lambda}_k(\boldsymbol{\ell}) \sum_{n\in\mathcal{N}} \boldsymbol{y}^k_n$. This follows since the dual is always feasible. Therefore, either the dual is unbounded in which case the primal is infeasible, i.e., $\Xi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\ell}) = \emptyset$, and the constraint is trivial. Else, by strong duality, the primal and dual must have the same objective value (zero). As a result, the constraints in Problem (A.11) associated with each $\boldsymbol{\ell} \in \partial\mathcal{L}$ can be written as

$$\tau \leq -\mathbf{e}^\top \boldsymbol{\theta}(\boldsymbol{\ell}) + \boldsymbol{b}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) - \sum_{\substack{k\in\mathcal{K}\\ \boldsymbol{\ell}_k\neq 0}} \left(\boldsymbol{y}^k_{\boldsymbol{\ell}_k} - 1\right) \boldsymbol{\nu}_k(\boldsymbol{\ell}) + \sum_{\substack{k\in\mathcal{K}\\ \boldsymbol{\ell}_k= 0}} \sum_{n\in\mathcal{N}} \boldsymbol{y}^k_n \boldsymbol{\beta}^k_n(\boldsymbol{\ell}) + \sum_{k\in\mathcal{K}} \boldsymbol{\lambda}_k(\boldsymbol{\ell}) \sum_{n\in\mathcal{N}} \boldsymbol{y}^k_n$$

$$\boldsymbol{\theta}_n(\boldsymbol{\ell}) \ \leq\ \boldsymbol{A}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) + \sum_{\substack{k\in\mathcal{K}\\ \boldsymbol{\ell}_k\neq 0}} \sum_{\nu\in\delta(\boldsymbol{\ell}_k)} \boldsymbol{x}_\nu \boldsymbol{\nu}_k(\boldsymbol{\ell}) - \sum_{\substack{k\in\mathcal{K}\\ \boldsymbol{\ell}_k= 0}} \sum_{\nu\in\delta(n)} \boldsymbol{x}_\nu \boldsymbol{\beta}^k_n(\boldsymbol{\ell}) \quad \forall n \in \mathcal{N}.$$

Finally, the last constraint in Problem (A.11) is satisfied if the linear program

$$\begin{aligned}
\min \quad & 0 \\
\text{s.t.} \quad & 0 \leq \boldsymbol{\xi}_n \leq 1 \quad \forall n \in \mathcal{N} \\
& \boldsymbol{A}\boldsymbol{\xi} \geq \boldsymbol{b} \\
& \boldsymbol{y}^k_{\boldsymbol{\ell}_k} \geq \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{\xi}_\nu \boldsymbol{x}_\nu + 1 \quad \forall k \in \mathcal{K} \,:\, \boldsymbol{\ell}_k \neq 0
\end{aligned}$$

is infeasible. Using strong duality, this occurs if the dual problem

$$\begin{aligned}
\max \quad & -\mathbf{e}^\top \boldsymbol{\theta}(\boldsymbol{l}) + \boldsymbol{\alpha}(\boldsymbol{\ell})^\top \boldsymbol{b} - \sum_{\substack{k \in \mathcal{K} \\ \boldsymbol{\ell}_k \neq 0}} \left(\boldsymbol{y}^k_{\boldsymbol{\ell}_k} - 1\right) \boldsymbol{\nu}_k(\boldsymbol{\ell}) \\
\text{s.t.} \quad & \boldsymbol{\theta}(\boldsymbol{\ell}) \in \mathbb{R}^N_+,\ \boldsymbol{\alpha}(\boldsymbol{\ell}) \in \mathbb{R}^R_+,\ \boldsymbol{\nu}(\boldsymbol{\ell}) \in \mathbb{R}^K_+ \\
& \boldsymbol{\theta}_n(\boldsymbol{\ell}) \leq \boldsymbol{A}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) + \sum_{\substack{k \in \mathcal{K} \\ \boldsymbol{\ell}_k \neq 0}} \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{x}_\nu \boldsymbol{\nu}_k(\boldsymbol{\ell}) \quad \forall n \in \mathcal{N}
\end{aligned}$$

is unbounded. Since the feasible region of the dual problem constitutes a cone, the dual problem is unbounded if and only if there is a feasible solution with an objective value of 1 or more. ■

A.5 Bender's Decomposition

We do not detail all the steps of the Bender's decomposition algorithm. We merely provide the initial relaxed master problem and the subproblems used to generate the cuts. We refer the reader to e.g., [37] for more details.

Relaxed Master Problem. Initially, the relaxed master problem only involves the binary variables of the Problem (1.5) and is expressible as

$$\max \left\{ \tau \,:\, \tau \in \mathbb{R},\, \boldsymbol{x} \in \mathcal{X},\, \boldsymbol{y}^1, \ldots, \boldsymbol{y}^K \in \mathcal{Y} \right\}.$$

Subproblems. As discussed in Section 1.5, Problem (1.5) decomposes by $\boldsymbol{\ell}$. Depending on the index $\boldsymbol{\ell}$ of the subproblem, there are two types of subproblems to consider. If $\boldsymbol{\ell} \in \mathcal{L}_0$, the subproblem is given by

$$\begin{aligned}
\min \quad & 0 \\
\text{s.t.} \quad & \boldsymbol{\theta}(\boldsymbol{\ell}),\, \boldsymbol{\beta}^k(\boldsymbol{\ell}) \in \mathbb{R}^N_+,\, \boldsymbol{\alpha}(\boldsymbol{\ell}) \in \mathbb{R}^R_+,\, \boldsymbol{\nu}(\boldsymbol{\ell}) \in \mathbb{R}^K_+,\, \boldsymbol{\lambda}(\boldsymbol{\ell}) \in \Delta_K(\boldsymbol{\ell}) \\
& \tau \;\leq\; -\mathbf{e}^\top \boldsymbol{\theta}(\boldsymbol{\ell}) + \boldsymbol{b}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) - \sum_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k \neq 0}} \left(\boldsymbol{y}^k_{\boldsymbol{\ell}_k} - 1 \right) \boldsymbol{\nu}_k(\boldsymbol{\ell}) + \ldots \\
& \qquad \ldots + \sum_{\substack{k \in \mathcal{K}: \\ \boldsymbol{\ell}_k = 0}} \sum_{n \in \mathcal{N}} \boldsymbol{y}^k_n \boldsymbol{\beta}^k_n(\boldsymbol{\ell}) + \sum_{k \in \mathcal{K}} \boldsymbol{\lambda}_k(\boldsymbol{\ell}) \sum_{n \in \mathcal{N}} \boldsymbol{y}^k_n
\end{aligned} \tag{$\mathcal{Z}_0(\boldsymbol{\ell})$}$$

$$\boldsymbol{\theta}_n(\boldsymbol{\ell}) \;\leq\; \boldsymbol{A}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) + \sum_{\substack{k \in \mathcal{K} \\ \boldsymbol{\ell}_k \neq 0}} \sum_{\nu \in \delta(l_k)} \boldsymbol{x}_\nu \boldsymbol{\nu}_k(\boldsymbol{\ell}) - \sum_{\substack{k \in \mathcal{K} \\ \boldsymbol{\ell}_k = 0}} \sum_{\nu \in \delta(n)} \boldsymbol{x}_\nu \boldsymbol{\beta}^k_n(\boldsymbol{\ell}) \;\; \forall n \in \mathcal{N}.$$

In a similar fashion, we define the subproblem associated with $\boldsymbol{\ell} \in \mathcal{L}_+$, given by

$$
\begin{aligned}
\min \quad & 0 \\
\text{s.t.} \quad & \boldsymbol{\theta}(\boldsymbol{\ell}) \in \mathbb{R}^N_+,\ \boldsymbol{\alpha}(\boldsymbol{l}) \in \mathbb{R}^R_+,\ \boldsymbol{\nu}(\boldsymbol{l}) \in \mathbb{R}^K_+ \\
& 1 \leq -\mathbf{e}^\top \boldsymbol{\theta}(\boldsymbol{l}) + \boldsymbol{b}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) - \sum_{\substack{k \in \mathcal{K} \\ \boldsymbol{\ell}_k \neq 0}} \left(\boldsymbol{y}^k_{\boldsymbol{\ell}_k} - 1 \right) \boldsymbol{\nu}_k(\boldsymbol{\ell}) \\
& \boldsymbol{\theta}_n(\boldsymbol{\ell}) \leq \boldsymbol{A}^\top \boldsymbol{\alpha}(\boldsymbol{\ell}) + \sum_{\substack{k \in \mathcal{K} \\ \boldsymbol{\ell}_k \neq 0}} \sum_{\nu \in \delta(\boldsymbol{\ell}_k)} \boldsymbol{x}_\nu \boldsymbol{\nu}_k(\boldsymbol{\ell}) \quad \forall n \in \mathcal{N}.
\end{aligned}
\tag{$\mathcal{Z}_+(\boldsymbol{\ell})$}
$$

Appendix B

Technical Appendix to Chapter 3

B.1 Omitted Proofs from Section 2.5.2

Proof of Proposition 4. Let $F : \mathbb{R}^N \to \mathbb{R}$ be an additive function in the form $F(\boldsymbol{u}) = \sum_{i=1}^{N} f(u_i)$ where $f : \mathbb{R} \to \mathbb{R}$ is a monotonically increasing and strictly concave function. We are focusing on group fairness where the utility of each individual is given by the average utility of their community. Hence, we can rewrite $F(\boldsymbol{u}) = \sum_{c \in \mathcal{C}} N_c f(u_c)$. Let $\boldsymbol{u} = \boldsymbol{u}(\mathcal{A})$ and $\boldsymbol{u}' = \boldsymbol{u}(\mathcal{A}')$ denote the utility vectors corresponding to neighboring solutions $\mathcal{A}$ and $\mathcal{A}'$, respectively. Suppose $\boldsymbol{u}$ and $\boldsymbol{u}'$ are sorted in ascending order and for all $c \in \mathcal{C}$, index c in both vectors corresponds to the same community, i.e., after the transfer the ordering of the utilities has not changed.

Furthermore, assume $\Sigma_{\kappa \in \mathcal{C}:\kappa \leq c} N_\kappa (u_\kappa - u'_\kappa) \geq 0$, $\forall c \in \mathcal{C}$ and $u_c > u'_c$ for some $c \in \mathcal{C}$. Clearly $\boldsymbol{u}$ and $\boldsymbol{u}'$ satisfy the assumptions of the influence transfer principle. We need to show that $\Sigma_{c \in \mathcal{C}} N_c f(u_c) > \Sigma_{c \in \mathcal{C}} N_c f(u'_c)$ or $\Sigma_{c \in \mathcal{C}} N_c \left(f(u_c) - f(u'_c)\right) > 0$.

The proof is by induction. We iteratively sweep the vectors $\boldsymbol{u}$ and $\boldsymbol{u}'$ from the smallest index to the largest and show that for any $\kappa \in \mathcal{C}$, $\Sigma_{c \leq \kappa} N_c \left(f(u_c) - f(u'_c)\right) \geq 0$ with inequality becoming strict for at least one κ. To do so we repeatedly use a property of strictly concave functions known as decreasing marginal returns. According to this property $f(x + \delta_x) - f(x) > f(y + \delta_y) - f(y)$ for $x < y$ and $\delta_x \geq \delta_y > 0$.

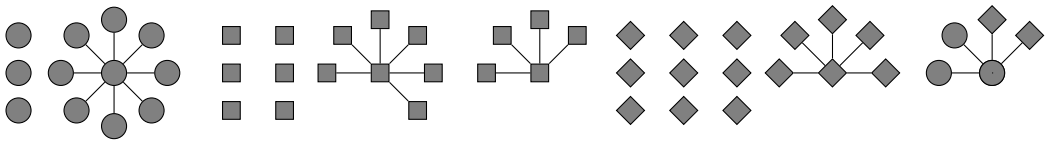


Figure B.1: An illustration for the graph used in the proof of Proposition 5 without the correct scaling. There are three communities (circle, square and diamond) and they all have size 100. The circle community consists of an "all-circle" star structure with 80 vertices, 14 isolated vertices and a mixed star structure (shared with the diamond community) with 6 circle vertices. The square community consists of two "all-square" star structures with sizes 60 and 10 plus a set of 30 isolated vertices. The diamond community consists of an "all-diamond" star structure with 30 vertices, 66 isolated vertices and a mixed star structure (shared with the circle community) with 4 diamond vertices.

More specifically, in our inductive step, we keep track of a "decrement budget" which we denote by Δ. Intuitively if we can show that $\Sigma_{c\leq\kappa}N_c\left(f(u_c)-f(u'_c)\right)>0$ with budget Δ for some κ, we can then use the decreasing marginal return property along with the assumption that $\boldsymbol{u}'$ is sorted to show that as long as $N_{\kappa+1}\left(u'_{\kappa+1}-u_{\kappa+1}\right)\leq\Delta$ it is the case that $\Sigma_{c\leq\kappa+1}N_c\left(f(u_c)-f(u'_c)\right)>0$. After each round we update the Δ and move on to the next element in the utility vectors.

Formally, let $\Delta=0$ to start at the begining of this inductive process. After visiting the cth community, we simply update Δ by $\Delta\leftarrow\Delta+N_c\left(u_c-u'_c\right)$. By the assumption of the transfer principle Δ is non-negative at all points of this iterative process and is strictly positive at some point during the process. Observe that $f(u_1)\geq f(u'_1)$ since $u_1\geq u'_1$ by the assumption of the transfer principle and monotonicity of f. We can use this as the base case. Since $\boldsymbol{u}$ and $\boldsymbol{u}'$ are sorted, given that Δ is non-negative, the fact that f is strictly concave (so that the decreasing marginal return property can be used) immediately implies that $\Sigma_{c\leq\kappa}N_c\left(f(u_c)-f(u'_c)\right)\geq 0$ at any iteration κ of the process. The inequality becomes strict for some κ given the assumption of the transfer principle. This proves the claim. ■

Proof of Proposition 5. Figure B.1 is an illustration of the graph that is used in the proof to witness the statement. We set $p=1$ (deterministic spread) and number of initial seeds $K=4$. Consider two choices of influencer vertices $\mathcal{A}$ and $\mathcal{A}'$. Let $\mathcal{A}$ denote the choice that consists

of the center of all-star structures that consist of a single community. Let $\mathcal{A}'$ denote the solution that is identical to $\mathcal{A}$ with the sole difference that only the center of one of the all-square structures is chosen and the last seed is selected to be the center of the star structure that is the mix of circle and diamond communities. Clearly these two solutions are neighboring. The average utilities for these solutions are (diamond $= 0.3$, square $= 0.7$, circle $= 0.8$) in $\boldsymbol{u}$ and (diamond $= 0.34$, square $= 0.6$, circle $= 0.86$) in $\boldsymbol{u}'$, respectively. Both solutions correspond to a total utility of 180 but the utility gap is $\Delta(\boldsymbol{u}) = 0.5$ for $\boldsymbol{u}$ as opposed to the utility gap of $\Delta(\boldsymbol{u}') = 0.52$ for $\boldsymbol{u}'$. So a welfare function that obeys the utility gap reduction should prefer $\boldsymbol{u}$ over $\boldsymbol{u}'$.

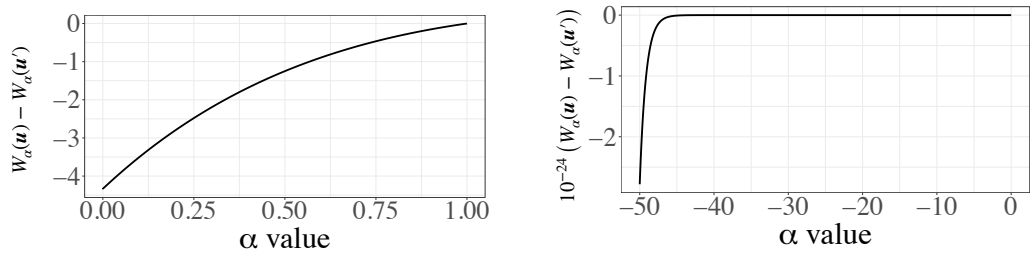


Figure B.2: The difference of $W_\alpha(\boldsymbol{u}) - W_\alpha(\boldsymbol{u}')$ on the vertical axis versus α on the horizontal axis for different welfare functions (this difference is scaled by a factor of 10^{-24} on the bottom panel). Top panel: $W_\alpha(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c u_c^\alpha/\alpha$ for $\alpha \in (0,1)$; bottom panel: $W_\alpha(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c u_c^\alpha/\alpha$ for $\alpha < 0$.

We now show that no welfare function that satisfies the first 5 principles will prefer $\boldsymbol{u}$ over $\boldsymbol{u}'$. Recall that such welfare functions are in the form $W_\alpha(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c u_c^\alpha/\alpha$ for $\alpha < 1$ and $\alpha \neq 0$, $W_\alpha(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c \log(u_c)$ for $\alpha = 0$. We verify this claim numerically. In particular Figure B.2 plots the difference of $W_\alpha(\boldsymbol{u}) - W_\alpha(\boldsymbol{u}')$ for $W_\alpha(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c u_c^\alpha/\alpha$ when $\alpha \in (0,1)$ (top panel) and $\alpha < 0$ (bottom panel). This difference is always negative so $\boldsymbol{u}'$ is preferred by these welfare functions. For $W_\alpha(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c \log(u_c), W_\alpha(\boldsymbol{u}) - W_\alpha(\boldsymbol{u}') \approx -4.3$.

We point out that the instance used in the proof (graph structure, probability of spread and the number of seeds) is designed with the sole purpose of simplifying the calculations of the utilities. It is possible to modify this instance to more complicated and realistic instances.

■

Proof of Proposition 6. Let $\mathcal{A}$ and $\mathcal{A}'$ denote two neighboring solutions with corresponding utility vectors $\boldsymbol{u} = \boldsymbol{u}(\mathcal{A})$ and $\boldsymbol{u}' = \boldsymbol{u}(\mathcal{A}')$. Let $\boldsymbol{u}$ denote any of the two utility vectors such that $\Sigma_{c\in\mathcal{C}} N_c u_c \geq \Sigma_{c\in\mathcal{C}} N_c u'_c$. Without loss of generality, we assume $\boldsymbol{u}'$ is sorted in ascending order of the utilities and $\boldsymbol{u}$ is permuted so that index $c \in \mathcal{C}$ in both $\boldsymbol{u}$ and $\boldsymbol{u}'$ corresponds to the same community. This is because we assume that W satisfies the symmetry principle due to which by permuting a utility vector the value of the welfare function does not change. Let ν and $\kappa \in \mathcal{C}$ denote the communities whose utilities are changed between $\boldsymbol{u}$ and $\boldsymbol{u}'$, i.e., we assume ν and κ are the two communities where taking influencer vertices from ν and giving them to κ will transfer $\boldsymbol{u}'$ into $\boldsymbol{u}$.

To satisfy the condition of the utility gap reduction principle, it should be the case that $u'_\nu \geq u'_\kappa$ (i.e., we transfer influencer vertices from the group with higher utility to a group with lower utility), otherwise after the transfer from $\boldsymbol{u}'$ to $\boldsymbol{u}$ the utility gap could not get smaller (i.e., $\Delta(\boldsymbol{u}) \geq \Delta(\boldsymbol{u}')$ in which case the utility gap reduction is not applicable).

Assuming $u'_\nu \geq u'_\kappa$, if $\Delta(\boldsymbol{u}) \geq \Delta(\boldsymbol{u}')$, again the assumption of the utility gap reduction principle is not satisfied, hence the principle is not applicable and there is no need to study this case. Therefore, we further assume $\Delta(\boldsymbol{u}) < \Delta(\boldsymbol{u}')$. We would like to show in this case a welfare function W that satisfies all the 5 other principles witnesses $W(\boldsymbol{u}) > W(\boldsymbol{u}')$.

By assumption $\Sigma_{c\in\mathcal{C}} N_c u_c \geq \Sigma_{c\in\mathcal{C}} N_c u'_c$. From this, it follows that:

$$\sum_{c\in\mathcal{C}} N_c \left(u_c - u'_c\right) \geq 0 \tag{B.1}$$

$$\Leftrightarrow N_\nu \left(u_\nu - u'_\nu\right) + N_\kappa \left(u_\kappa - u'_\kappa\right) \geq 0 \tag{B.2}$$

$$\Leftrightarrow \sum_{y\in\mathcal{C}: y\leq x} N_y (u_y - u'_y) \geq 0,\ \forall x \in \mathcal{C}, \tag{B.3}$$

where both inequalities (B.2) and (B.3) follow directly from the fact that the utilities of all the other communities are the same in both $\boldsymbol{u}$ and $\boldsymbol{u}'$. Finally, since $u_\kappa > u'_\kappa$ (we are transferring influencer vertices to the community κ), we can apply the influence transfer principle to show that $W(\boldsymbol{u}) > W(\boldsymbol{u}')$ as claimed. ■

Proof of Lemma 3. As we have shown earlier welfare functions $W_\alpha(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c \log(u_c)$ for $\alpha = 0$ and $W_\alpha(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c u_c^\alpha/\alpha$ for $\alpha < 1, \alpha \neq 0$ satisfy all the first 5 principles. In [117], the authors show that the composition of a non-decreasing concave function (in our case $\log(x), \alpha = 0$ or x^α/α for $\alpha < 1, \alpha \neq 0$) and a non-decreasing submodular function (in our case $u_c(\mathcal{A})$) is submodular. Since the sum of submodular functions is submodular, our proposed class of welfare functions is submodular. Our welfare functions also satisfy monotonicity. This is because $u_c(\mathcal{A})$ is monotonically non-decreasing so its composition with another monotonically non-decreasing function ($\log(x)$ for $\alpha = 0$ or x^α/α for $\alpha < 1, \alpha \neq 0$) will be monotonically non-decreasing. Since our welfare functions are the sum of monotonically non-decreasing function they are also monotone. ■

B.2 Leximin Fairness and Social Welfare

In this section, we show that leximin fairness can be captured by our welfare maximizing framework. See [86] for more details.

Proposition 13. *Welfare optimization is equivalent to the leximin fairness, i.e., there exists a constant α_0, such for $\alpha < \alpha_0$, an optimal solution to the welfare maximization satisfies leximin fairness and vice versa.*

Proof. Let $\overline{\boldsymbol{u}} = (\overline{u}_1, \ldots, \overline{u}_N) \succcurlyeq \boldsymbol{u}(\mathcal{A})\ \forall \mathcal{A} \in \mathcal{A}^\star$, where "$\succcurlyeq$" is the lexicographic ordering sign and it indicates that $\overline{\boldsymbol{u}}$ is a leximin fair solution (w.l.o.g. and with a slight abuse of notation, we assume that both $\overline{\boldsymbol{u}}$ and $\boldsymbol{u}(\mathcal{A})$ are sorted in increasing order). We aim to show that $\exists \alpha_0 < 0$ such that

for any $\alpha \leq \alpha_0$, $\Sigma_{i=1}^{N} \overline{u}_i^{\alpha}/\alpha \geq \Sigma_{i=1}^{N} u_i^{\alpha}(\mathcal{A})/\alpha,\ \forall \mathcal{A} \in \mathcal{A}^{\star}$. For simplicity we multiply both sides of the inequality by $-1/\alpha$ and since $\alpha < 0$ the direction of the inequality sign does not change.

We now prove this inequality by contradiction. Suppose $\forall \alpha_1 < 0$, $\exists \alpha < \alpha_1$ such that $\Sigma_{i=1}^{N} - \overline{u}_i^{\alpha} < \Sigma_{i=1}^{N} - u_i^{\alpha}(\mathcal{A}), \exists \mathcal{A} \in \mathcal{A}^{\star}$. Since $\overline{\boldsymbol{u}}$ is a leximin solution then by definition $\overline{u}_1 \geq u_1(\mathcal{A})$. We consider two cases. First suppose $\overline{u}_1 > u_1(\mathcal{A})$.

$$\sum_{i=1}^{N} -\overline{u}_i^{\alpha} < \sum_{i=1}^{N} -u_i^{\alpha}(\mathcal{A}) \Leftrightarrow$$
$$\frac{\sum_{i=1}^{N} -\overline{u}_i^{\alpha}}{\min(\overline{u}_1, u_1(\mathcal{A}))^{\alpha}} < \frac{\sum_{i=1}^{N} -u_i^{\alpha}(\mathcal{A})}{\min(\overline{u}_1, u_1(\mathcal{A}))^{\alpha}} =$$
$$\frac{\sum_{i=1}^{N} -\overline{u}_i^{\alpha}}{u_1(\mathcal{A})^{\alpha}} < \frac{\sum_{i=1}^{N} -u_i^{\alpha}(\mathcal{A})}{u_1(\mathcal{A})^{\alpha}} \Rightarrow$$
$$\lim_{\alpha \to -\infty} \frac{\sum_{i=1}^{N} -\overline{u}_i^{\alpha}}{u_1(\mathcal{A})^{\alpha}} \leq \lim_{\alpha \to -\infty} \frac{\sum_{i=1}^{N} -u_i^{\alpha}(\mathcal{A})}{u_1(\mathcal{A})^{\alpha}} \Rightarrow$$
$$0 \leq -N_1.$$

This is a contradiction since $N_1 > 0$. Now, suppose $\overline{u}_1 = u_1(A)$. In this case, we can eliminate the first terms that involve $\overline{u}_1^{\alpha}$ and u_1^{α} from the two sides of inequality and redo the above steps iteratively starting from the second biggest element in $\overline{u}^{\alpha}$.

Next, we prove the other direction. Let us assume $\overline{\boldsymbol{u}}$ is a utility vector such that $\exists \alpha_0 < 0, \forall \alpha \leq \alpha_0$, $\sum_{i=1}^{N} -\overline{u}_i^{\alpha} \geq \sum_{i=1}^{N} -u_i^{\alpha}(\mathcal{A}),\ \forall \mathcal{A} \in \mathcal{A}^{\star}$. W.l.o.g, we can assume that $\overline{u}_1 \neq u_1$ otherwise we can remove those terms that are equal and the proof still holds. However, we assume this for ease of exposition. It follows that

$$\frac{\sum_{i=1}^{N} -\overline{u}_i^{\alpha}}{\min(\overline{u}_1, u_1)^{\alpha}} \geq \frac{\sum_{i=1}^{N} -u_i^{\alpha}(\mathcal{A})}{\min(\overline{u}_1, u_1)^{\alpha}},\ \forall \mathcal{A} \in \mathcal{A}^{\star}.$$

If $\min(\overline{u}_1, u_1) = \overline{u}_1$ meaning that $u_1 > \overline{u}_1$ we have $-C - \epsilon(\alpha) \geq -\delta(\alpha, \mathcal{A}),\ \forall \mathcal{A} \in \mathcal{A}^{\star}$ where $C > 0$ is a constant (equal to the number of entities in $\overline{\boldsymbol{u}}$ that are equal to $\overline{u}_1$) and both $\epsilon \geq 0$ and $\delta \geq 0$ are functions of α and can be made arbitrarily small by decreasing α. This is a contradiction

which means that $\min(\overline{u}_1, u_1) = u_1$, i.e., $\overline{u}_1 \geq u_1$. By continuing this procedure, we can establish that $\overline{\boldsymbol{u}} \succcurlyeq \boldsymbol{u}$. This completes the proof. ■

B.3 Omitted Proofs from Table 2.1

In this section we provide detailed description of the entries of Table 2.1 and their derivations.

B.3.1 Monotonicity

Proposition 14. *Exact DP does not satisfy monotonicity.*

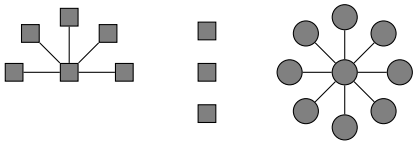


Figure B.3: Companion figure to Proposition 14. The network consists of two communities circle and square, each of size N.

Proof. Let $K = 2$ and $p \in (0, 1)$. Consider a graph $\mathcal{G}$ as shown in Figure B.3 consisting of two communities, square and circle, each of size N (for large enough N). The circle community consists of a star network of size N. The square community contains a star network of size $2 + p(N-2)$ and $(N-2)(1-p)$ singletons. Consider two solutions $\mathcal{A}$ and $\mathcal{A}'$. $\mathcal{A}$ will select a seed from the periphery of the star for the circle community and allocate the other seed to the center of the star for the square community. $\mathcal{A}'$ on the other hand allocates each of the seeds to the center of the stars. Let $\boldsymbol{u} = \boldsymbol{u}(\mathcal{A})$ and $\boldsymbol{u}' = \boldsymbol{u}(\mathcal{A}')$ denote the corresponding allocations of $\mathcal{A}$ and $\mathcal{A}'$. The utility vectors for these allocations are $\boldsymbol{u} = ((1 + p + p^2(N-2))/N, (1 + p + p^2(N-2))/N)$ and $\boldsymbol{u}' = ((1 + p(N-1))/N, (1 + p + p^2(N-2))/N)$, respectively. Clearly, $\boldsymbol{u} < \boldsymbol{u}'$. So by monotonicity $\boldsymbol{u}'$ is preferred to $\boldsymbol{u}$. However, only $\boldsymbol{u}$ satisfies the exact DP. Hence, DP does not satisfy monotonicity. ■

Proposition 15. *Approximate DP does not satisfy monotonicity.*

Proof. Consider a graph $\mathcal{G}$ as shown in Figure B.4 consisting of two communities, square and circle, each of size N. We choose an arbitrary $\delta \in (0,1)$, to reflect the arbitrary strictness of a decision maker. Let $\delta < p < \sqrt{\delta}$, $K = 2$ and $N > \max\left(3p/(p-\delta), 1/(\delta - p^2)\right)$. The optimal solution $\mathcal{A}$ of the influence maximization problem chooses the center of the star and any disconnected square vertex. In $\mathcal{A}$, the utility of circle and square communities are $(1 + (N-1)p)/N$ and $(1+2p)/N$, respectively and the utility gap exceeds δ (so this solution does not satisfy the DP constraints). By imposing DP, any fair solution is to choose one vertex from the periphery of the circle community and one from the isolated square vertices. For a fair solution $\mathcal{A}'$, the utilities of circle and square are $\left(1 + p + p^2(N-2)\right)/N$ and $(1+2p^2)/N$, respectively. Given the range of N, the utility gap is less than δ so approximate DP is satisfied. Since the utility of both communities have degraded, any monotone welfare function will prefer $\mathcal{A}'$ (and its corresponding utility vector) over $\mathcal{A}$. However, only $\mathcal{A}'$ is DP fair and hence it is preferred over $\mathcal{A}$ by DP. We point out that the graph used in the proof is directed. This is for ease of exposition. It is possible to create a more complex example with an undirected graph. ■

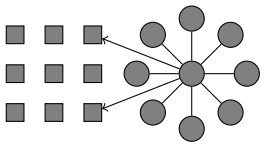


Figure B.4: Companion figure to Proposition 15. The network consists of two communities circle and square, each of size N. All edges except the two shown by arrows are undirected meaning that influence can spread both ways.

Proposition 16. *Consider a general fairness notion as a set of constraints in the form of* $\mathcal{F} = \left\{\boldsymbol{u} \in [0,1]^C : u_c \geq l_c, \ \ \forall c \in \mathcal{C}\right\}$ *where* $l_c \ \forall c \in \mathcal{C}$ *are arbitrary lower-bound values. The considered fairness notion satisfies the monotonicity principle.*

Proof. Let $\mathcal{A}$ and $\mathcal{A}' \in \mathcal{A}^\star$ denote two solutions whose corresponding utility vectors $\boldsymbol{u} = \boldsymbol{u}(\mathcal{A})$ and $\boldsymbol{u}' = \boldsymbol{u}(\mathcal{A})$ are feasible ($\boldsymbol{u}, \boldsymbol{u}' \in \mathcal{F}$) such that $\boldsymbol{u} < \boldsymbol{u}'$. Given the objective function of the influence maximization is equivalent to $\Sigma_{c \in \mathcal{C}} N_c u_c(\mathcal{A})$ and that all N_c values are positive the

objective values of $\boldsymbol{u}'$ is strictly better than $\boldsymbol{u}$. Hence, $W(\boldsymbol{u}) < W(\boldsymbol{u}')$ and the monotonicity is satisfied. ■

As we have shown in Section 2.4, both maximin and DC can be written as constraints that are compatible with the fairness definition in Proposition 16. The utilitarian solution corresponds to setting all the lower bounds to 0. Hence all of them satisfy monotonicity.

Corollary 1. *DC, MMF and utilitarian satisfy monotonicity.*

B.3.2 Symmetry

It is straightforward to show that DP (exact and approximate), maximin and utilitarian fairness satisfy the symmetry principle. DC, however, does not satisfy the symmetry principle. Based on its definition, DC can place different lower-bounds on the utility of different communities. Hence, by permuting a utility vector we may no longer be able to satisfy the DC constraints (see Definition 2.2).

B.3.3 Independence of Unconcerned Individuals

Proposition 17. *Exact and approximate DP do not satisfy the independence of unconcerned individuals.*

Proof. Consider two utility vectors: $\boldsymbol{u} = ((1+3\delta)/8, (1-\delta)/8)$ and $\boldsymbol{u}' = ((1+\delta)/4, (1-\delta)/8)$ for $\delta \in [0, 1)$. Both exact and approximate DP strictly prefer $\boldsymbol{u}$ over $\boldsymbol{u}'$. Let us substitute the second component of both vectors by $(1+\delta)/4$. Therefore, we obtain $\boldsymbol{v} = \boldsymbol{u}|^2(1+\delta)/4 = ((1+3\delta)/8, (1+\delta)/4)$ and $\boldsymbol{v}' = \boldsymbol{u}'|^2(1+\delta)/4 = ((1+\delta)/4, (1+\delta)/4)$. In contrast to the previous case, both approximate and exact DP prefer $\boldsymbol{v}'$ over $\boldsymbol{v}$. Note that while the construction does does not involve an instance of the influence maximization problem, it is possible to provide such an instance to witness the claim as follows.

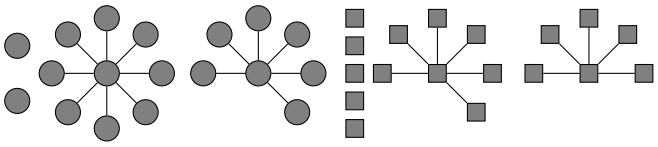


Figure B.5: Companion figure to Proposition 17. The network consists of two communities circle and square each of size N.

Figure B.5 demonstrates the instance witnessing the claim. We consider an influence maximization problem with two communities: circle (first component of the utility vector) and square (second component of the utility vector), each of size N (for N large enough). We assume $p = 1$ and $K = 2$. The circle community consists of three components: two star components of size $N(1+\delta)/4$ (small) and $N(1+3\delta)/8$ (large) and $5N(1-\delta)/8$ isolated vertices. The square community consists of three components as well: two star components of size $N(1-\delta)/8$ (small) and $N(1+\delta)/4$ (large) and $N(5-\delta)/8$ isolated vertices. Solution $\boldsymbol{u}$ ($\boldsymbol{u}'$) corresponds to selecting a seed vertex from the large (small) star component of the circle community and a vertex from the small star component of the square community. Allocation $\boldsymbol{v}$ ($\boldsymbol{v}'$) corresponds to selecting a seed vertex from the large (small) star component of the circle community a vertex from the large star component of the square community. Note that the choice of $p = 1$ is merely for the ease of exposition and the example network can be modified to accommodate $p < 1$. ■

Henceforth, we only discuss utility vectors when appropriate. In all such cases, there exist instances of the influence maximization problem which witness these utility vectors. We have demonstrated one such instance in the proof of Proposition 17, but we omit the details from the remaining proofs for simplicity.

Proposition 18. *DC does not satisfy the independence of unconcerned individuals.*

Proof. Consider an instance of the influence maximization problem with 2 communities were the lower bound set by DC for both communities is 0.4. Also consider 2 solutions with corresponding utility vectors $\boldsymbol{u} = (0.5, 0.5)$ and $\boldsymbol{u}' = (0.5, 0.3)$. Therefore, only $\boldsymbol{u}$ satisfies DC and hence DC prefers $\boldsymbol{u}$ over $\boldsymbol{u}'$. Let us substitute the first component with 0.35. Therefore, we obtain

$\boldsymbol{v} = \boldsymbol{u}|^1 0.35 = (0.35, 0.5)$ and $\boldsymbol{v}' = \boldsymbol{u}'|^1 0.35 = (0.35, 0.3)$. In contrast to the previous case, both solutions are infeasible with respect to the DC (hence -∞ welfare, see Section 2.5.4). Therefore, while $W(\boldsymbol{u}) > W(\boldsymbol{u}')$, it does not hold that $W(\boldsymbol{v}) > W(\boldsymbol{v}')$. ■

Proposition 19. *MMF does not satisfy the independence of unconcerned individuals.*

Proof. Consider an instance of the influence maximization problem with 3 communities of equal size. Also consider 2 solutions with corresponding utility vectors $\boldsymbol{u} = (0.3, 0.6, 0.4)$ and $\boldsymbol{u}' = (0.3, 0.2, 0.8)$. Maximin fairness strictly prefers $\boldsymbol{u}$ over $\boldsymbol{u}'$. Let us substitute the first component with 0.1. Therefore, we obtain $\boldsymbol{v} = \boldsymbol{u}|^1 0.1 = (0.1, 0.6, 0.4)$ and $\boldsymbol{v}' = \boldsymbol{u}'|^1 0.1 = (0.1, 0.2, 0.8)$. Maximin fairness is indifferent between $\boldsymbol{v}$ and $\boldsymbol{v}'$ (both have the same worst-case utility and total utility) which shows that maximin fairness does not satisfy the independence of unconcerned individuals. ■

Note that the utilitarian satisfies the independence of unconcerned individuals because if $W(\boldsymbol{u}) = \Sigma_{c\in\mathcal{C}} N_c u_c > \Sigma_{c\in\mathcal{C}} N_c u'_c = W(\boldsymbol{u}')$ then $W(\boldsymbol{u}|^c b') < W(\boldsymbol{u}'|^c b')$ since N_c, u_c, u'_c and b' are all non-negative.

B.3.4 Affine Invariance

Exact DP satisfies affine invariance principle because a linear transformation over a uniform vector will remain uniform. However, for approximate DP this is not the case. More particularly, for any utility vector $\boldsymbol{u}$ that is δ-DP for $\delta \in (0, 1)$ and an affine transformation of the form $\boldsymbol{u}' = \alpha\boldsymbol{u} + \beta$, $\boldsymbol{u}'$ satisfies $\alpha\delta$-DP. Therefore, for $\alpha > 1$, $\boldsymbol{u}'$ does not satisfy δ-DP. Similarly, DC does not satisfy this principle either. This is because after the transformation the constraints may not be satisfied (e.g., when $\alpha < 1/\min_{c\in\mathcal{C}} U_c$). It is known that MMF satisfies this principle [33]. The same holds for the utilitarian objective because if $\Sigma_{c\in\mathcal{C}} N_c u_c > \Sigma_{c\in\mathcal{C}} N_c u'_c$, then $\Sigma_{c\in\mathcal{C}} N_c \alpha u_c + \beta > \Sigma_{c\in\mathcal{C}} N_c \alpha u'_c + \beta$ since $\alpha > 0$.

B.3.5 Influence Transfer Principle

Proposition 20. *Exact and approximate DP do not satisfy the influence transfer principle.*

Proof. Let $\delta \in [0, 1)$ denote the parameter of DP. Consider utility vectors $\boldsymbol{u} = ((1+\delta)/2, 0)$ and $\boldsymbol{u}' = (\delta, 0)$. The sizes of the all the communities are the same. Based on the influence transfer principle, $W(\boldsymbol{u}) > W(\boldsymbol{u}')$, however, DP strictly prefers $\boldsymbol{u}'$ over $\boldsymbol{u}$. ■

Proposition 21. *DC does not satisfy the influence transfer principle.*

Proof. Let $\boldsymbol{u} = (0.5, 0.5)$ and $\boldsymbol{u}' = (0.3, 0.6)$ denote the utility vectors of two allocations where sizes of the all the communities are the same. Suppose the lower bounds set by DC are 0.25 and 0.55, respectively. This means that only $\boldsymbol{u}'$ satisfies DC. Based on the transfer principle, $W(\boldsymbol{u}) > W(\boldsymbol{u}')$, however, $\boldsymbol{u}$ does not satisfy DC and DC strictly prefers $\boldsymbol{u}'$ over $\boldsymbol{u}$. ■

Proposition 22. *MMF does not satisfy the influence transfer principle.*

Proof. Consider two utility vectors $\boldsymbol{u} = (0.2, 0.4, 0.6)$ and $\boldsymbol{u}' = (0.2, 0.2, 0.8)$. The sizes of the all the communities are the same. A fairness notion satisfying the influence transfer principle strictly prefers $\boldsymbol{u}$ over $\boldsymbol{u}'$. However, Maximin is indifferent between $\boldsymbol{u}$ and $\boldsymbol{u}'$ as they both obtain the same worst-case utility. ■

Proposition 23. *Utilitarian does not satisfy the influence transfer principle.*

Proof. Consider two utility vectors $\boldsymbol{u} = (0.5, 0.5)$ and $\boldsymbol{u}' = (0.3, 0.7)$. The sizes of the all the communities are the same. Based on the transfer principle, $W(\boldsymbol{u}) > W(\boldsymbol{u}')$, however, the utilitarian approach is indifferent between $\boldsymbol{u}$ and $\boldsymbol{u}'$ as both solutions lead to the same total utility. ■

B.3.6 Utility Gap Reduction

Proposition 24. DP *satisfies the utility gap reduction if and only if* $\delta = 0$.

Proof. It is easy to show that if $\delta = 0$, DP satisfies the utility gap reduction principle. We can prove this by contradiction. Suppose that $\delta = 0$ and DP does not satisfy the utility gap reduction principle. From this, it follows that given two utility vectors $\boldsymbol{u}, \boldsymbol{u}'$ such that $\sum_{c \in \mathcal{C}} N_c u_c \geq \sum_{c \in \mathcal{C}} N_c u'_c$ if $\Delta \boldsymbol{u} < \Delta \boldsymbol{u}'$, DP can strictly prefer $\boldsymbol{u}'$. If $\boldsymbol{u}'$ is preferred, then $\boldsymbol{u}'$ is feasible and it must be that $\Delta \boldsymbol{u}' = 0$ and $\Delta \boldsymbol{u} < 0$ which is not possible. Next, we show that if $\delta \neq 0$, DP does not satisfy this principle over all instances of the influence maximization problem.

For the proof, we use the example used in the proof of Proposition 5. In that setting there are two solutions with utility gap 0.52 and 0.5 with the same total utility. First, let us assume $\delta > 0.02$. In this case, since both solutions satisfy DP constraints, they are both feasible and the total utility of both solutions is equal ($= 180$), however, DP does not *strictly* prefer the solution with smaller gap. In fact, both solutions are feasible with the same objective value and DP does not favor one solution to the other. For $\delta \leq 0.02$, we can use the same example graph (see Figure B.1) and add enough isolated vertices to each community until the gap between the solutions becomes small enough to pass the δ threshold. ■

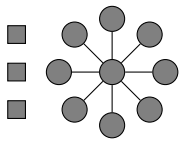


Figure B.6: Companion figure to Proposition 25 of a graph with two communities: N black vertices and $N/3$ white vertices for $N = 9$. We choose $K = 4$ and arbitrary $p < 1$. All edges are undirected, meaning that influence can spread both ways.

Proposition 25. DC *does not satisfy utility gap reduction principle.*

Proof. Consider the network $\mathcal{G}$ as in Figure B.6 consisting of two communities white and black with size $N/3$ and N, respectively. Suppose $K = 4$ and $p < 1$. Without DC, an optimal solution places one seed vertex at the center of the black group and allocates the remaining 3 vertices to the white group. We let $\boldsymbol{u}$ denote this solution. Thus, the utility of the black and white groups will be equal to $(1 + (N-1)p)/N$ and $9/N$. Since there is no edge between the black and white

communities, DC (See Definition 2.2) reduces to how to optimally choose one seed vertex from white and the remaining 3 from the black group. After imposing DC, the utility of the black and white groups will be equal to $(3+(N-3)p)/N$ and $3/N$. We let $\boldsymbol{w}$ denote one such solution that satisfies DC. While $\boldsymbol{u}$ has a higher total utility and a smaller utility gap, DC strictly prefers $\boldsymbol{w}$ with higher utility gap and lower total utility. Therefore, DC does not satisfy utility gap reduction principle. ■

Proposition 26. *MMF does not satisfy the utility gap reduction*

Proof. We prove the statement via the example in Figure B.7 which depicts a network with three groups: blue, black and white. We fix $K = 1$ and $p > 3/4$. The graph corresponds to the case where $p = 1$ but the example will hold for arbitrary p by setting the number of isolated green vertices to be $\lceil 21/p \rceil$.

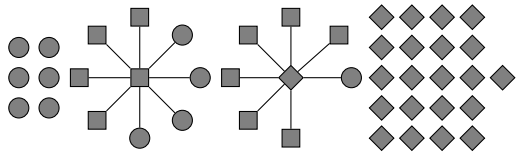


Figure B.7: Companion figure to Proposition 26 for the case of $p = 1$. The network consists of three groups: white, blue and black. The edges are undirected so the influence can spread both ways. For arbitrary p, the number of isolated black vertices should scale to $\lceil 21/p \rceil$.

Consider one solution that targets the center of the bigger star component. Thus, the utilities of blue, black and white will be $(1 + 4p)/11$, $4p/11$ and 0, respectively. This results in utility gap equal to $(1 + 4p)/11$. By imposing leximin, the optimal fair solution selects the center of the smaller star component and the optimal fair utilities of blue, black and white will be $6p/11$, $p/11$ and $1/(1 + \lceil 21/p \rceil)$, respectively and we observe a utility gap $6p/11 - 1/(1 + \lceil 21/p \rceil) \geq 6p/11 - 1/22 > (1 + 4p)/11$, where we used $p > 3/4$ in the last inequality. In conclusion, while the first solution has a higher total utility ($= 9$) and lower utility gap compared to the second solution (total utility $= 8$), leximin still strictly prefers the second solution. This concludes the proof. ■

Proposition 27. *Utilitarian does not satisfy the utility gap reduction*

Proof. Consider an instance of the influence maximization problem where all the communities are of size 10. Let $\boldsymbol{u} = (0.5, 0.5, 0.5)$ and $\boldsymbol{u}' = (0.2, 0.5, 0.8)$. Both $\boldsymbol{u}$ and $\boldsymbol{u}'$ achieve the same total utility (=15). Thus, the utilitarian approach is indifferent between $\boldsymbol{u}$ and $\boldsymbol{u}'$. According to the utility gap reduction $\boldsymbol{u}$ is strictly preferred. We note that in this special instance, both influence transfer principle and utility gap reduction apply and according to both principles $\boldsymbol{u}$ is strictly preferred to $\boldsymbol{u}'$. ■

B.4 Omitted Details from Section 2.6

B.4.1 Estimating the SBM Parameters for Landslide Risk Management

In order to qualitatively and formatively describe the network structure, the research team conducted several in-person semi-structured interviews in Sitka, Alaska from 2018-2020. These interviews were conducted with individuals who were identified as "community leaders" or "popular community members" through word-of-mouth, and then subsequently through respondent-driven sampling, a broader range of community members were interviewed (n=14). In these semi-structured interviews, respondents were asked to 1) sort and describe community groups and 2) identify "cliques" and "isolates" as they relate to an early landslide warning system. The former resulted in developing, to the extent possible, discrete a priori community groups. The latter helped to inform the relationships between and within these groups. The interviewer took notes which listed the responses and through a tallying and pile sorting exercise, attempted to seek consensus in definitions of community groups. The formative research resulted in cliques based on occupation, political affiliation, age, and local recreational activities. Many cliques were overlapping with shared attributes (e.g. people from two different occupations share a political affiliation and frequent the same local pub), however for the purposes of this formative exercise, these community groups were qualitatively coerced into discrete classifications. These resulted

in 16 community groups that include political affiliation, time spent in Sitka (e.g. new arrival or tourist vs. long-term resident), occupation, and whether or not a parent of a child in the public-school system. The community size estimates were developed based on a 2018 Sitka Economic Development Survey, particularly for the occupation-based community groups, as well as publicly available voter records for political affiliations. Several attributes, namely age, specific occupation, time spent in Sitka, and parental status were unavailable in existing datasets, and therefore required the use of proxies and assumptions for estimating community group sizes. Once the community group sizes were estimated, based on the formative research notes on social cohesion, cliques, and isolates, we further developed assumptions on within and between-community connectedness. For example, if a respondent suggested that there may be very close relationships between two cliques, we assumed a higher relative p(b) than between two cliques which had less similar attributes. For simplicity, we limited the absolute probabilities for withing-community and between-community probabilities between 0.00 and 0.10. We then sense-checked these absolute probabilities with several of the initial formative research respondents. These absolute probabilities were then organized into a 16×16 adjacency matrix to facilitate simulations for influence maximization.

B.4.2 Relative Community Sizes

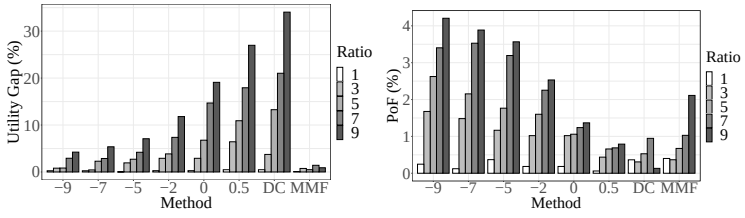


Figure B.8: Utility gap and PoF for various relative community sizes where the ratio changes from 1 to 9.

In this section we study the effect of relative community size on both utility gap and efficiency. We consider synthetic samples of SBM network consisting of two communities each of size 100 and

we gradually increase the size of one community from 100 to 900 in order to study its effect on the utility values of each community. We set $q_c = 0.005$ and $q_{cc'} = 0.001$. Results are summarized in Figure B.8. This result indicates that the utility gap increases with the relative community size, suggesting that minorities can be adversely affected without appropriate fairness measures in place. We also note that the strength of our approach is in its flexibility to trade-off fairness with efficiency. We may encounter scenarios where the fairness-efficiency trade-offs are mild (as in the particular setting of Figure B.8), but this does not undermine our approach as there are many practical situations (as discussed in real-world applications in the paper) where clearly this is not the case and our approach can handle all those cases effectively. DC exhibits relatively high utility gap. This is because by definition DC allocates more resources to communities that "will do better with the resources" and it does not always show an aversion to inequality, a result which we show theoretically in Section 2.5.4.

B.4.3 Suicide Prevention Application

Network Name	# of Vertices	# of Edges	White	Black	Hispanic	Mixed Race	Other
W1MFP	219	217	16.4	41.5	20.5	16.4	5.0
W2MFP	243	214	16.8	36.6	21.8	22.2	2.4
W3MFP	296	326	22.6	34.4	15.2	22.9	4.7
W2SPY	133	225	55.6	10.5	–	22.5	11.3
W3SPY	144	227	63.0	–	–	16.0	20.0
W4SPY	124	111	54.0	16.1	–	14.5	15.3

Table B.1: Racial composition (%) after pre-processing as well as the number of vertices and edges of the social networks [17]

Influence maximization has been previously implemented for health promoting interventions among homeless youth [188, 185]. In this section, we consider the problem of training a set of individuals who can share information to help prevent suicide (e.g., how to identify warning signs of suicide among others). We present simulation results over six different social networks of homeless youth from a major city in US as described in [17]. We provide aggregate summaries of these networks (e.g., size, edge density and community statistics) in Table B.1. The data set

Measure (%)	K	Fairness Approach							
		$\alpha = -5$	$\alpha = -2$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 0.9$	DC	Maximin	IM
utility gap	5	4.8	7.5	8.5	9.7	11.4	8.2	**3.5**	12.5
	10	4.6	6.6	7.3	9.5	12.9	6.9	**2.0**	11.7
	15	3.6	5.2	5.9	8.9	13.5	7.6	**2.4**	15.3
	20	3.6	4.4	5.9	7.3	14.0	5.8	**2.3**	17.2
	25	2.6	3.5	4.6	5.8	13.2	7.0	**2.0**	16.6
	30	2.4	3.2	4.3	6.4	8.6	8.2	**2.0**	15.7
PoF	5	6.9	5.1	3.4	1.5	**0.7**	14.4	16.6	0.0
	10	6.6	3.8	2.8	0.7	**0.1**	14.3	13.1	0.0
	15	3.8	2.5	1.6	1.1	**0.1**	14.6	10.5	0.0
	20	4.6	3.8	2.9	2.0	**1.0**	13.9	10.5	0.0
	25	4.0	3.2	2.5	1.9	**1.0**	13.4	9.9	0.0
	30	3.9	3.4	2.9	2.3	**1.8**	12.7	10.8	0.0

Table B.2: Summary of the utility gap and PoF results averaged over 6 different real world social networks for various budget, fairness approaches and baselines. Numbers in bold highlight the best values in each setting (row) across different approaches.

consists of six different social networks of homeless youth from a major city in US as described in detail in [17]. Each social network consists of 7 racial groups, namely, *Black or African American, Latino or Hispanic, White, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander and Mixed race.* Each individual belongs to a single racial group. We use these partitioning by race to define our communities. However, to avoid misinterpretation of the results, we combine racial groups with a population $< 10\%$ of the network size N under the "Other" category. After this pre-processing step, each dataset will contain 3 to 5 communities. Results are summarized in Table B.1. We remark that the absent of a racial category in a given network is due to their small sizes and hence being merged into the "Other" category after pre-processing (e.g., Hispanic in network W2SPY.)

We compare our welfare-based framework for different values of α against DC, MMF and influence maximization without fairness considerations (IM). Table B.2 provides a summary of the results averaged over all network instances where the numbers in bold highlight the best values (minimum utility gap and PoF) for each budget and across different fairness approaches. As seen, IM typically has a large utility gap (up to 17.2% for $K = 20$ which is significant because the total influence is only 28.40%). By imposing fairness we can reduce this gap. In fact, we observe

that across different values of α ranging from -5 to 0.5, there is a decreasing trend in utility gap, where for $K = 20$ and with $\alpha = -5$, we are able to decrease the utility gap by 3.6%. Consistent with previous results on SBM networks, both MMF and $\alpha = -5$ exhibit very low utility gaps, however, MMF results in higher PoF. Furthermore, across the range of α we observe a mild trade-off between fairness and utility. This shows that in these networks enforcing fairness comes at a low cost, though as we see in the landslide setting, this is not always the case.

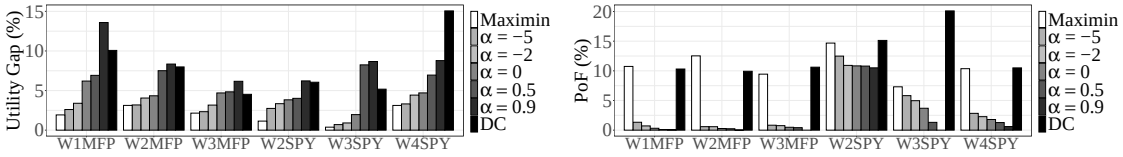


Figure B.9: Top and bottom panels: utility gap and PoF for each real world network instances ($K = 30$).

Figure B.9 shows the results for each network separately (X axis) for a fixed budget $K = 30$. Figure B.9 shows that the trade-offs can also be very network-dependent (compare e.g. W2SPY and W3MFP). This highlights the crucial need for a flexible framework that can be easily adjusted to meaningfully compare these trade-offs.

Appendix C

Technical Appendix to Chapter 4

C.1 Proof of Proposition 8

Proof. We let the $\boldsymbol{X}_q = 1$ be the event where $P(\boldsymbol{X}) = q$ ($\boldsymbol{X}_q = 0$ otherwise). It holds that:

$$
\begin{aligned}
V(\pi_M) &= \mathbb{E}\left[\sum_{r\in\mathcal{R}} \pi(r \mid \boldsymbol{X}) Y(r)\right] \\
&= \sum_{q\in\mathcal{Q}} \mathbb{P}(\boldsymbol{X}_q = 1)\, \mathbb{E}\left[\sum_{r\in\mathcal{R}} \pi(r \mid \boldsymbol{X}) Y(r) \,\middle|\, \boldsymbol{X}_q = 1\right] \\
&= \sum_{q\in\mathcal{Q}} \mathbb{P}(\boldsymbol{X}_q = 1)\, \mathbb{E}\left[\sum_{r\in\mathcal{R}} \pi(r \mid \boldsymbol{X}) \left(Y(r) - Y(0)\right) \,\middle|\, \boldsymbol{X}_q = 1\right] + \\
&\qquad \mathbb{P}(\boldsymbol{X}_q = 1)\, \mathbb{E}\left[\sum_{r\in\mathcal{R}} \pi(r \mid \boldsymbol{X}) Y(0) \,\middle|\, \boldsymbol{X}_q = 1\right] \\
&= \sum_{q\in\mathcal{Q}} \mathbb{P}(\boldsymbol{X}_q = 1)\, \mathbb{E}\left[\sum_{r\in\mathcal{R}} \pi(r \mid \boldsymbol{X}) \left(Y(r) - Y(0)\right) \,\middle|\, \boldsymbol{X}_q = 1\right] + \mathbb{P}(\boldsymbol{X}_q = 1)\, \mathbb{E}\left[Y(0) \mid \boldsymbol{X}_q = 1\right] \\
&= \sum_{q\in\mathcal{Q}} \mathbb{P}(\boldsymbol{X}_q = 1)\, \mathbb{E}\left[\sum_{r\in\mathcal{R}} \pi(r \mid \boldsymbol{X}) \left(Y(r) - Y(0)\right) \,\middle|\, \boldsymbol{X}_q = 1\right] + C \\
&= \sum_{q\in\mathcal{Q}} \mathbb{P}(\boldsymbol{X}_q = 1) \sum_{r\in\mathcal{R}} \pi(r \mid \boldsymbol{X})\, \mathbb{E}\left[Y(r) - Y(0) \mid \boldsymbol{X}_q = 1\right] + C \\
&= \sum_{q\in\mathcal{Q}} \frac{\lambda_q}{\lambda_{\mathcal{Q}}} \sum_{r\in\mathcal{R}} \frac{f_{qr}}{\lambda_q} \tau_{qr} + C \\
&= \sum_{q\in\mathcal{Q}} \sum_{r\in\mathcal{R}} \frac{f_{qr}\tau_{qr}}{\lambda_{\mathcal{Q}}} + C,
\end{aligned}
$$

where $C = \sum_{q \in \mathcal{Q}} \mathbb{P}(\boldsymbol{X}_q = 1)\, \mathbb{E}\left[Y(0) \mid \boldsymbol{X}_q = 1\right] = \mathbb{E}\left[Y(0)\right]$. ■

C.2 Proof of Proposition 9

Proof. We first prove part one and show the conditional independence for each component Y_r of the potential outcome vector. The proof is in the same vein as the balancing scores in the causal inference literature which is essentially a low-dimensional summary of the feature space that facilitates causal inference for observational data in settings with many features. For binary potential outcomes, we have

$$\begin{aligned}
\mathbb{P}(Y_r = 1 \mid \boldsymbol{S}, R) &= \mathbb{E}[Y_r \mid \boldsymbol{S}, R] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_r \mid \boldsymbol{S}, R, \boldsymbol{X}\right] \mid \boldsymbol{S}, R\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_r \mid \boldsymbol{S}, \boldsymbol{X}\right] \mid \boldsymbol{S}, R\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_r \mid \boldsymbol{X}\right] \mid \boldsymbol{S}, R\right] \\
&= \mathbb{E}\left[S_r \mid \boldsymbol{S}, R\right] \\
&= S_r,
\end{aligned}$$

where the third line follows the assumption of the proposition and the fourth line holds since $\boldsymbol{S}$ is essentially a function of $\boldsymbol{X}$ and can be dropped. We also show

$$\begin{aligned}
\mathbb{P}(Y_r = 1 \mid \boldsymbol{S}) &= \mathbb{E}[Y_r \mid \boldsymbol{S}] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_r \mid \boldsymbol{S}, \boldsymbol{X}\right] \mid \boldsymbol{S}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_r \mid \boldsymbol{X}\right] \mid \boldsymbol{S}\right] \\
&= \mathbb{E}\left[S_r \mid \boldsymbol{S}\right] \\
&= S_r.
\end{aligned}$$

We proved $\mathbb{P}(Y_r = 1 \mid \boldsymbol{S}, R) = \mathbb{P}(Y_r = 1 \mid \boldsymbol{S})$. We now prove the second part of the proposition.

$$\mathbb{P}\left(\mathbb{P}(R = r \mid \boldsymbol{X} = \boldsymbol{x}) > 0\right) = 1 \Rightarrow \mathbb{P}\left(\mathbb{P}(R = r, \boldsymbol{X} = \boldsymbol{x}) > 0\right) = 1$$

$$\begin{aligned}\mathbb{P}\left(\mathbb{P}(R = r, \boldsymbol{X} = \boldsymbol{x}) > 0\right) &= \mathbb{P}\left(\mathbb{P}(R = r, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{S} = \boldsymbol{s}) > 0\right) \\ &\leq \mathbb{P}\left(\mathbb{P}(R = r, \boldsymbol{S} = \boldsymbol{s}) > 0\right).\end{aligned}$$

It follows that $\mathbb{P}\left(\mathbb{P}(R = r, \boldsymbol{S} = \boldsymbol{s}) > 0\right) = 1$ for all values of $\boldsymbol{s}$. ■

C.3 Computational Results

HMIS Data Preparation. We used HMIS dataset collected between 2015 and 2017 across 16 communities in the United States. The dataset contains 10,922 homeless youth and 3464 PSH and RRH resources combined. We removed all those with veteran status (54 data points), pending and unknown outcomes (4713 data points). We grouped Hawaiian/Pacific Islander, Native American, Hispanics, Asian under 'Other' category as no significant statistical inference can be made on small set of observations within each individual category. Further, we removed 6 data points with no gender information. We use a median date 08/13/2015 as the cut-off date to separate train and test sets.

Outcome Estimation. Figure C.1 depicts the average outcome across different score values $\mathbb{E}[Y(r) \mid S = s]\ \forall r \in \mathcal{R}$, using the DR estimate. Under SO, after $S = 8$, there is a significant drop in average outcome. Average outcomes under PSH and RRH also exhibit a decline with score. However, they remain highly effective even for high-scoring youth.

Propensity Score. In order to evaluate different policies using IPW and DR methods, we estimated the propensity scores, i.e., $\pi_0(R = r \mid \boldsymbol{X} = \boldsymbol{x})$. Table C.1 summarizes the accuracy

	Model	In-Sample Accuracy (%)	Out-of-Sample Accuracy (%)
NST Score	Multinomial Regression	72.5	73.7
	Neural Network	76.4	76.5
	Decision Tree	76.3	76.2
	Random Forest	76.4	76.3
All Features	Multinomial Regression	75.4	73.5
	Neural Network	80.4	77.2
	Decision Tree	79.2	78.5
	Random Forest	99.7	79.3

Table C.1: Prediction accuracy for propensity estimation using HMIS data.

across different models. We consider two models, one that uses only the NST score and one that uses the entire set of features in the data. We observe that, even though the policy recommendations only use NST score, including other features help improve the accuracy. In addition, the decision tree and random forest are the top-performing models. Although random forest exhibits over-fitting (in-sample accuracy = 99.6%) its out-of-sample accuracy (79.3%) outperforms other models. In addition to accuracy, the propensity models should be well-calibrated. That is, the observed probability should match the predicted probability. We plot the reliability diagrams in Figure C.2, where $y-$axis is the observed probability in the data and the $x-$axis is the predicted value. The dots correspond to values of different bins. A well-calibrated model should lie on the $y = x$ diagonal line.

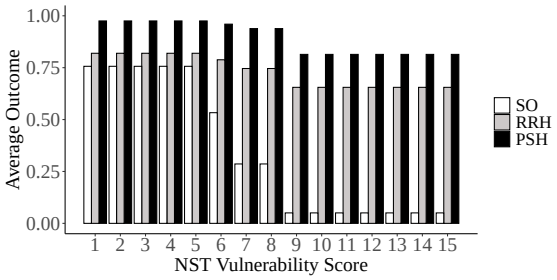


Figure C.1: Probability of exiting homelessness across the NST score range estimated using the DR method.

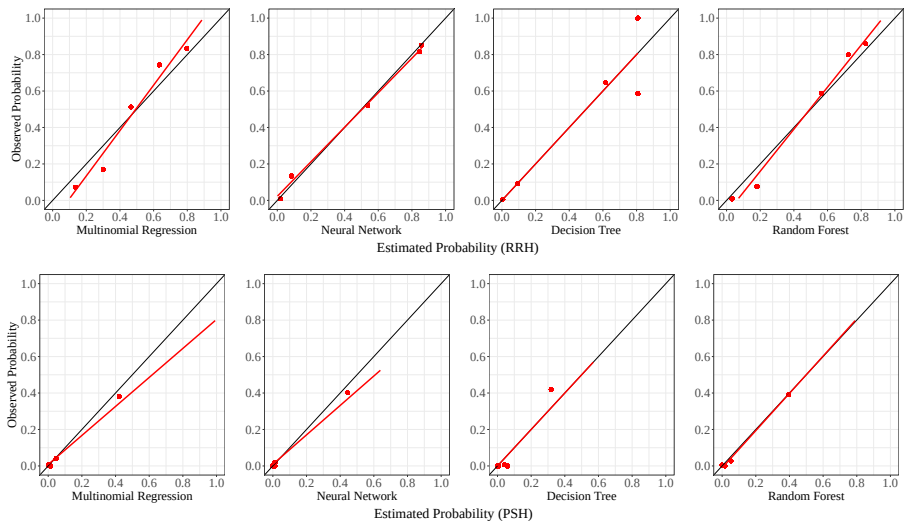


Figure C.2: Reliability diagram of propensity estimation, RRH (top) and PSH (bottom).

As seen in Figure C.2, random forest and neural network models have relatively better calibration property. Finally, in our model selection, we take fairness considerations into account. In particular, we study the calibration of the models across different demographic groups for which fair treatment is important. Since ultimately we use the probability estimates, not the binary prediction, it is important to ensure that across different demographic groups, the models are well-calibrated. We adopted test-fairness notion [51]. We fit a model to predict the resource one receives, based on the predicted propensities and demographic features. In a well-calibrated model across demographic groups, the coefficients of the demographic attributes should not be statistically significant in the prediction. For the predicted values of the random forest model none of the demographic attributes coefficients were found to be statistically significant. In addition, the model were calibrated within groups with coefficient near 1. Regression results are summarized in Table C.2. Hence, we chose random forest as the model of historical policy π_0.

Outcome Estimation. In the direct method, one estimates the (counterfactual) outcomes under different resources by fitting the regression models $\mathbb{P}(Y \mid \boldsymbol{X} = \boldsymbol{x}, R = r) \ \forall r \in \mathcal{R}$. For model selection, we followed the same procedure as propensity score estimation. Table C.3 summarizes the accuracy of different models for each type of resource.

Coeffs.	Estimates	p-value
Intercept	-0.012	0.066
PSH pred.	**0.985**	<2e-16
Race = 2	0.011	0.204
Race = 3	0.002	0.803
Gender = 2	-0.006	0.469
Age = 2	0.006	0.485

Coeffs.	Estimates	p-value
Intercept	-0.054	5.7e-05
RRH pred.	**1.125**	< 2e-16
Race = 2	-0.007	0.586
Race = 3	-0.014	0.394
Gender = 2	0.000	0.987
Age = 2	-0.003	0.813

Table C.2: Propensity calibration within group for PSH (left) and RRH (right) of random forest model. None of the coefficients of the demographic attributes are found to be significant. In addition, the coefficient associated with the predicted probability is close to 1 in both models, suggesting that the model is well-calibrated even when we control for the demographic attributes.

	Model	PSH	RRH	SO
NST	Logistic Regression	83.1	78.8	90.0
	Neural Network	83.9	78.9	90.0
	Decision Tree	83.9	78.9	90.0
	Random Forest	83.1	78.6	90.0
NST + Demographic	Logistic Regression	83.1	78.8	90.0
	Neural Network	81.6	78.3	90.3
	Decision Tree	83.9	78.8	90.0
	Random Forest	83.9	78.1	90.0
All Features	Logistic Regression	81.9	82.2	90.3
	Neural Network	83.9	78.8	86.8
	Decision Tree	74.3	81.1	90.0
	Random Forest	83.9	81.4	90.0

Table C.3: Out-of-Sample Accuracy (%) of different outcome estimation models (outcome definition in Figure 3.4).

Considering the reliability diagrams in Figure C.3, we observe that logistic regression models are well-calibrated across different resources. We also investigated test-fairness of logistic regression where we fit the observed outcome against the predicted outcome and demographic features. Results are summarized in Table C.4. As seen, the coefficients of demographic features are not significant, suggesting that test-fairness is satisfied.

Optimal Matching Topology for Fairness over Age.

Figure C.4 depicts the policies when fairness over age is imposed. According to this figure, across all score values youth below 17 years are eligible for PSH. On the other hand, mid- and

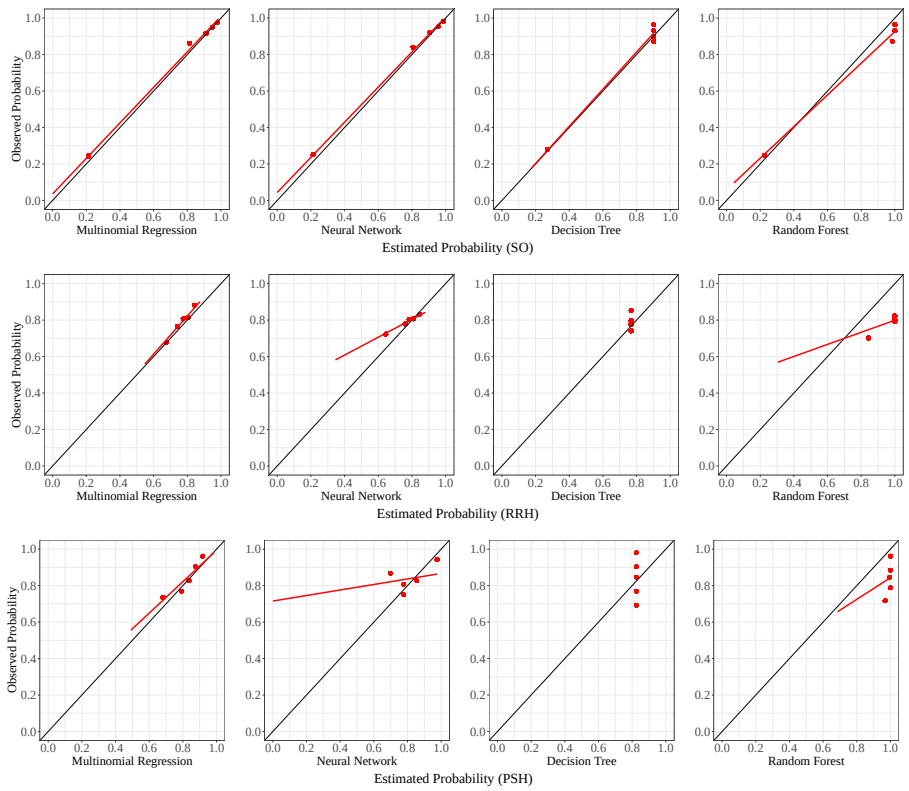


Figure C.3: Reliability diagram of outcome, SO (top), RRH (middle) and PSH (bottom).

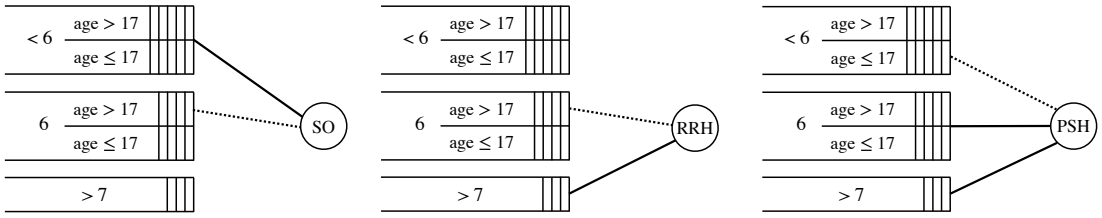


Figure C.4: The matching topology split by resource type: left (SO), middle (RRH) and right (PSH). The solid line indicates that the resource is connected to the entire queue. The dotted line indicates connection to a sub-group within the queue. For example, in the left figure, SO is only connected to the individuals with NST = 6 and age over 17.

high-scoring youth over 17 years old, are eligible for PSH. We further imposed constraints to ensure within each score group, the connections are the same for different age groups. Figure C.5 illustrates the resulting matching topology, according to which individuals who score above 7 are eligible for RRH and PSH, regardless of their age. Those who score 6 are eligible for all three resource types. Finally, All youth with score below 6 are only eligible for SO. We observe that all

	Coeffs.	Estimates	p-value
PSH	Intercept	0.147	0.489
	PSH pred.	**0.853**	0.000
	Race = 2	-0.021	0.666
	Race = 3	-0.061	0.324
	Gender = 2	0.003	0.954
	Age = 2	0.079	0.202
RRH	Intercept	-0.122	0.645
	RRH pred.	**1.172**	0.000
	Race = 2	0.028	0.386
	Race = 3	0.025	0.504
	Gender = 2	-0.021	0.433
	Age = 2	0.003	0.931
SO	Intercept	0.035	0.148
	SO pred.	**0.974**	<2e-16
	Race = 2	-0.000	0.973
	Race = 3	0.023	0.226
	Gender = 2	-0.008	0.618
	Age = 2	-0.011	0.542

Table C.4: Outcome calibration of logistic regression model within group under PSH, RRH and SO. None of the coefficients of the demographic attributes are found to be significant. In addition, the coefficient associated with the predicted probability is close to 1 in both models, suggesting that the model is well-calibrated even when we control for the demographic attributes.

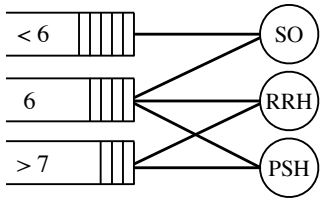


Figure C.5: Fair topology (age)

individuals who belong to a certain queue, regardless of their age, are eligible for the same types of resources. As a result of combining the queues that depended on age, the worst-case policy value across the age groups decreased from 0.74 to 0.69 which still outperforms the SQ (data) with worst-case performance of 0.64.