CARNEGIE MELLON UNIVERSITY

# The Role of Noise, Proxies, and Dynamics in Algorithmic Fairness

Nil-Jana Akpinar

June 2023

Department of Statistics and Data Science
& Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee**

Alexandra Chouldechova, Chair
Zachary Lipton, Chair
Hoda Heidari
Arun Kuchibhotla
Cyrus DiCiccio

*Dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics and Machine Learning.*

*To my grandparents, Ingrid and Dilaver.*

## Acknowledgments

I would like to express my deepest gratitude and appreciation to everyone who has made this PhD journey possible. I consider myself incredibly fortunate to have met so many wonderful researchers and friends on this journey.

First and foremost, I would like to extend my heartfelt thanks to my advisor Alexandra Chouldechova for her guidance, consistent support, and countless insightful research discussions. I am deeply grateful to you for always showing up for me in a compassionate but realistic manner and you have quickly become an important role model. I want to thank my advisor Zachary Lipton for his invaluable research inputs. Thank you for always seeing the bigger picture, and starting every conversation unrelated to research.

I am grateful to all the researchers who have shaped my path and research interests. Thank you to my committee members Hoda Heidari, Arun Kuchibhotla, and Cyrus DiCiccio for many productive research conversations. Thank you to Umut Acar and Aaditya Ramdas for getting me started with research at CMU. Thank you to Stefan Feuerriegel and Sören Bartels for supporting me when I first discovered my interest in Statistics and Machine Learning. Thank you to my dear friend Sören Künzel for sparking this interest and encouraging me to apply to a PhD at CMU. None of this would have been possible without you.

This acknowledgement section would not be complete without sincerely thanking all of my friends and family who have made these last five years an incredible journey filled with joy, support and unwavering encouragement. Thank you to my classmate, office mate, roommate and most importantly friend Mikaela Meyer who has been my ride or die since day 1 of the PhD journey. Thank you to my friend Holly Bossart for always having an open ear. Thank you to my pandemic pod Mikaela, Holly, Nic, Sasha, and Nick for keeping me sane in incredibly difficult times. Thank you to my fellow PhD students and friends including but not limited to Alec, Beomjo, Brendan, Catherine, Charvi, Emily, Emma, Feyza, Gabrielle, Ian, Jinjin, Julia, Leqi, Meg, Nari, Nupoor, Octavio, Roger, Sid, Theresa, and Tudor for making this PhD a great experience all around. Thank you to Nora, Martin, and all my family and friends back home for always believing in me.

# Abstract

Machine learning is increasingly used to aid or automate decision making. Yet, algorithmic solutions often suffer from bias and disparate impact across demographic groups. For many application settings, the mechanisms by which bias arises and the effects of applying fairness-aware learning methods are not sufficiently understood. In this thesis, I focus on differential noise and missingness as drivers of bias (Part I), and long-term dynamics of fairness promoting interventions (Part II).

Part I of the thesis presents three studies on the impacts of differentially missing observations, differential feature mismeasurement, and differentially informative proxies. First, we discuss how geographical differences in victim crime reporting rates can lead to outcome disparities in predictive policing systems. Second, we explore the fairness implications of differential feature under-reporting in the setting of public sector risk assessment instruments, and propose technical solution approaches. The third study proposes a sandbox tool to evaluate fairness-enhancing algorithms under different types of artificially injected bias. Potential use cases for the tool are demonstrated via case studies.

Part II of the thesis comprises a study of long-term dynamics of fairness intervention in connection recommendation. Using both simulation and theoretical limit analysis, we demonstrate how typical fairness-promoting interventions can fail to promote equity in second order variables of interest such as network sizes.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Machine Learning is increasingly used to aid or automate decision making in public and private sector domains such as policing, health care, criminal justice or recommendation. While these technologies have the potential to improve the quality and efficiency of decisions, many algorithmic solutions have been shown to have disparate impact across demographic groups (Angwin et al., 2016; Chouldechova et al., 2018a), which can lead to long-lasting adverse effects for already vulnerable populations (Buolamwini and Gebru, 2018; Bolukbasi et al., 2016; Sweeney, 2013). There is a growing recognition that meaningfully addressing issues of algorithmic bias requires a multi-disciplinary effort that involves regulators, practitioners, civil society, and researchers from across the computational sciences, social sciences, and humanities. In recent years, the machine learning literature has proposed many general-purpose mathematical definitions of fairness and algorithms to enforce them (Chouldechova and Roth, 2020; Dwork et al., 2012; Hardt et al., 2016; Agarwal et al., 2018a; Zhang et al., 2018). Despite these advancements, for many application settings, the mechanisms by which bias arises and the effects of applying fairness-aware learning methods in deployed algorithms are not sufficiently well understood.

My work brings to bear methodologies at the interface of statistics and machine learning to examine

(i) differential noise and missing data as drivers of bias in predictive policing and risk assessment, and

(ii) the long-term effects of fairness-promoting interventions in dynamic recommendation settings.

Specifically, this thesis takes a data- and problem-centered perspective to study the role of noise, proxies and dynamics in algorithmic fairness.

Data bias is a key driver of outcome disparities across many application domains. In Part I of this dissertation, I focus on data bias in the form of differential noise and missingness. This can take the form of (i) differentially missing observations (Chapter 2 & 4), (ii) differential feature mismeasurement (Chapter 3 & 4), or (iii) label noise in the form of differentially informative proxies (Chapter 2 & 4).

Chapter 2 focuses on a predictive policing case study that highlights the effects of proxy outcomes and differential missingness of observations. Our analysis is based on a simulation patterned after district-level victimization and crime reporting survey data for Bogotá, Colombia, and uses spatio-temporal point process modeling similar to commercial predictive policing software (Mohler et al., 2011; 2015). We find that differences in victim crime reporting rates can lead to geographical outcome disparities in common crime hot spot prediction algorithms by systematically over-estimating crime prevalence in certain regions while underestimating it in others. This bias occurs even if no arrest data or other data from police initiated contact is used. We further show that reweighting observed data using survey-based victim crime reporting rates generally does not mitigate the bias. These findings supplement previous work on fairness in predictive policing which has predominantly focused on dynamical aspects, i.e. the dangerous feedback loops arising from models based on discovered crime data (Lum and Isaac, 2016; Ensign et al., 2018b). Relying on data obtained from victim reports rather than from police initiated contacts may prevent vicious cycles, however our study suggests that it does not necessarily lead to equitable outcomes. This showcases the importance of considering both the role of dynamics and the role of proxies when auditing algorithmic systems for fairness.

In Chapter 3, we study the fairness impact of differential feature under-reporting. Here, we use the term ‘under-reporting’ to refer to a specific type of data missingness that is prevalent in administrative data settings. Predictive risk models in the public sector are generally developed using administrative data that is more complete for individuals who have more greatly relied on publicly-funded services. For example, in the U.S., administrative records may contain medical claim information for individuals covered by Medicaid and Medicare but not for those privately insured. Thus, when a count or indicator of 0 is observed in the data, it is not known whether the feature is correctly observed as 0 (e.g. the individual in fact had 0 emergency room visits in the past year) or the feature is under-reported (e.g. the individual had 3 ER visits, but was privately insured, so no record of the visits exists in the local government’s administrative data). We propose and study an analytically tractable model of differential feature under-reporting to characterize the impact of this type of data bias on algorithmic fairness. The results demonstrate that, theoretically, differential under-reporting can lead to either increasing and decreasing disparities. However, our experiments on semi-synthetic and real-world data suggest that the case of decreasing disparities rarely occurs in practice. We also demonstrate that popular missing data methods like omission of mismeasured features or data imputation, in general, do not lead to more equitable outcomes as compared to using the mismeasured data directly. Instead, we propose and evaluate a method based on loss augmentation and imputation at prediction time to combat the bias introduced through under-reporting.

In Chapter 4, I propose a sandbox tool to evaluate fairness-related interventions under different types of artificially injected bias. The biases here can be understood as differential noise and missingess in features and labels, as well as differential missingness in observations, and are injected into otherwise (assumed to be) bias-free data sets. The bias injection idea can be used to assess the effectiveness of algorithmic remedies in the presence of specific bias types in a controlled environment. In particular, the framework allows one to test whether a particular fairness-enhancing method can successfully alleviate a specific type of bias by comparing the predictions after intervention to the labels before bias injection. Rather than placing the focus solely on the fairness-enhancing algorithm, the proposed framework considers a pair of bias-type and algorithm to obtain a more holistic picture and ensure the algorithm in fact alleviates the cause of unfairness. We demonstrate the utility of the sandbox tool via proof-of concept case studies using synthetic data sets, various types of biases (incl. under-representation bias, sampling bias, label bias, feature mismeasurement, and confounding bias), and in-and post-processing intervention methods enforcing various fairness metrics (incl. Equalized Odds, Equality of Opportunity, and Demographic Parity).

In Part II of this thesis, we turn our attention towards dynamics of fairness and fairness-promoting interventions over time. Chapter 5 presents a study of long-term dynamics of fairness intervention in Connection Recommendation. Connection recommendation is a key component of many social media platforms and, in some cases, accounts for more than 50% of the social network graph (LinkedIn). Despite the dynamical nature of recommender systems, most previous work on ranking and recommendation fairness assesses the efficacy of fairness intervention by evaluating a fixed fairness criterion — usually a parity condition in recommendation lists — through the lens of a one-shot static setting (Patro et al., 2022). This ignores potential effects of the intervention on the recommendation dynamics over time. Through a simulation framework, we demonstrate how recommendation patterned after the systems employed by

web-scale social networks promotes a group-wise rich-get-richer phenomenon and, although seemingly fair in aggregate, common exposure and utility parity interventions fail to mitigate bias amplification in average network size in the long term. We present a theoretical limit analysis assuming a stylized connection recommendation system based on Pólya urns to precisely characterize why the interventions are not sufficient. We conclude that reaching a stable fair equilibrium of network sizes requires a more in-depth consideration of intervention dynamics and measures of recommendation slate fairness that are suitable proxy targets.

# Part I

# Differential noise and missing data

# Chapter 2

# Fairness implications of unreported crime in predictive policing

Based on (Akpinar et al., 2021): Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT 2021).

## 2.1 Introduction

Police departments around the world have been experimenting with computer-aided place-based predictive policing systems for over two decades. In a 1998 National Institute of Justice survey, 36% of police agencies employing over 100 sworn officers reported having the computing capability and data infrastructure to digitally generate crime maps (Mamalian and La Vigne, 1999). Just a few years later, over 70% of agencies reported using such maps to identify crime hot spots as part of a broader adoption of CompStat approaches to policing (Weisburd et al., 2008). More modern incarnations of predictive policing date back to 2008, when the Los Angeles Police Department (LAPD) began its explorations of these systems, followed shortly thereafter by efforts such as the New York Police Department's use of tools developed by firms including Azavea, KeyStats and PredPol (2012+). Far from being a US-centric phenomenon, such systems are widely used throughout Europe, the UK, and China (Jansen, 2018; Babuta and Oswald, 2020;

Sprick, 2020).

More recently, predictive policing systems have come under scrutiny due to their lack of transparency (Winston, 2018) and concerns that they may lead to further over-policing of minority communities by virtue of being trained on biased or "dirty" data (Lum and Isaac, 2016; Ensign et al., 2018b; Richardson et al., 2019). Critics commonly point to the possibility that such systems may produce dangerous feedback loops, vicious cycles wherein data on recent arrests is used to deploy police in still greater numbers to neighbourhoods where they zealously seek out suspicious activity and conduct even more arrests. Recent work by Lum and Isaac (2016) and Ensign et al. (2018b) has demonstrated both empirically and theoretically how such feedback loops can arise.

Proponents and developers of predictive policing technologies have argued that such analyses are based on models of crime and policing that do not accurately reflect the types of data used as inputs to such systems, nor the types of crime that they seek to predict. The analysis of Lum and Isaac (2016), for instance, convincingly demonstrates how using data on drug arrests in Oakland, CA as inputs to a self-exciting point process (SEPP) model of the kind used in PredPol would result in high concentrations of policing in racial and ethnic minority neighbourhoods. Yet PredPol has stated that they do not use data on drug-related offenses (or traffic citations) in generating their predictions, nor do they use data on arrests (PredPol, 2017a). Azavea, the creators and former owners of the HunchLab product, likewise note that their models focus on property and violent crimes, and the crime data they use is based on victim reporting rather than arrests (Cheetham, 2019).

Secondly, proponents and developers have argued that prior studies incorrectly assume that targeted policing strategies lead to an escalation in crime detection and, correspondingly, arrests. However, the adoption of hot spot policing strategies is predicated on an anticipated *deterrence* effect. Studies of the impacts of predictive policing on property and violent crimes and on arrests at targeted locations have produced mixed results. A 2014 analysis of a randomized controlled experiment (RCT) conducted by RAND in Shreveport, Louisiana found no statistical evidence of crime reduction in the prediction-targeted locations compared to control locations (Hunt et al., 2014). Another RCT conducted in Pittsburgh reported a 34% drop in serious violent crime in "temporary hot spots" and a 24% drop in "chronic hot spots" (Fitzpatrick et al., 2018). This study found no evidence of crime displacement to nearby locations, and reported that a total of 4 arrests took place during the experiment's 20,000 hot spot patrols. A peer-reviewed study published by researchers affiliated with PredPol concluded that, while arrests were higher at predicted locations, they were lower or comparable once the counts were adjusted for differences in crime rate (Brantingham et al., 2018). PredPol has reported crime drops ranging from 8-30% depending on the jurisdiction and type of crime (PredPol, 2017b). While none of these counterarguments establish (or even claim) that the victim crime reporting data used to inform predictive policing systems is free from bias or leads to unbiased practices, they do point to a need for further investigation in settings that more closely mirror standard practice. Our work presents an initial step in this direction.

In this chapter we empirically demonstrate how predictive policing systems trained exclusively on victim crime reporting data (rather than arrest data) may nevertheless suffer from significant biases due to variation in reporting rates. Our analysis is based on a simplified crime simulation patterned after district-level crime statistics for Bogotá, Colombia released by Bogotá's chamber of commerce, Cámara de Comercio de Bogotá (CCB). We demonstrate that variation in crime reporting rates can lead to significant mis-allocation of police. These findings corroborate the effects of differential victim crime reporting on predictive policing models hypothesized in (Cuellar and De-Arteaga, 2020). We also discuss the limitations of using reporting rates from existing crime victimization surveys to attempt to correct for such biases. The code for this chapter is available at `https://github.com/nakpinar/diff-crime-reporting.git`.

## 2.2 Background and related work

### 2.2.1 Feedback loops and other biases in predictive policing

Having already described the work of Lum and Isaac (2016), we focus here on (Ensign et al., 2018b). Ensign et al. (2018b) theoretically characterize why feedback loops occur by modeling arrest-based predictive policing systems via a generalized Pólya urn model. Their analysis also considers a scenario in which both reported and detected crimes (i.e., arrests) are used to update beliefs about existing crime rates. In the latter case they show that if the reported crime rates are an accurate reflection of underlying crime, then feedback loops can be avoided if either (a) underlying crime rates across regions are uniform to begin with; or (b) detected crimes aren't considered at all. As we will discuss, there is considerable variation in the extent to which reported crimes reflect true underlying crime levels.

Richardson et al. (2019) present three case studies where there is evidence that “dirty data” may have biased the targets of predictive policing systems. Their case studies focus primarily on person-based predictive policing systems. In the case of Maricopa County, Arizona, however, the authors report one instance in which biased data may have informed a PredPol system used by the Mesa Police Department and an RTMDx system used by the Glendale Police Department. As the authors note, due to the lack of transparency surrounding what data was used and how, it is difficult to draw definitive conclusions. This, however, does not make the documented patterns of biased practices against Maricopa County’s Latino residents any less concerning.

### 2.2.2 Victim crime reporting

Many countries and local governments conduct crime victimization surveys to better understand factors that drive differences in crime reporting rates, and to assess discrepancies between official crime statistics and victimization-based measures of criminal activity. According to the 2018 report released by the Bureau of Justice Statistics, which oversees the annual US National Crime Victimization Survey (NCVS), 61% of aggravated assaults, 63% of robberies, 38% of simple assaults, and only 25% rapes/sexual assaults are reported to police (Morgan and Oudekerk, 2019). In this section we briefly overview different sources of disparities in victim crime reporting in the US context. We note that, while our data simulation is based on a 2014 survey conducted in Bogotá — and crime reporting rates are observed to be considerably lower there — a number of our conclusions apply to geography-associated disparities in reporting rates in general. In particular, our analysis indicates that, to the extent that these sources of disparity coincide with geography, we can expect significant under- or over-targeting to result.

The likelihood that a crime is reported to police has been found to be greater for older victims (Hashima and Finkelhor, 1999; Watkins, 2005; Bosick et al., 2012; Baumer, 2002) and when the victim is a woman (Baumer and Lauritsen, 2010). It is also greater if a third party is present (Baumer and Lauritsen, 2010), if a weapon is present or the victim is injured (Baumer and Lauritsen, 2010; Xie et al., 2006). Furthermore, reporting rates tend to increase with the degree to which the victim is of higher socioeconomic status than the offender, which in part accounts for the greater likelihood of white victims reporting crimes perpetrated by black offenders for crimes such as assaults (Xie and Lauritsen, 2012). However, Xie and Lauritsen (2012) also observe that black-on-black assaults had by far the highest reporting rate in their study (44%, compared to 25-33% for other racial pairs). This finding of high reporting rates for intraracial black-on-black crimes was also reported in (Avakame et al., 1999). In other words, while some might expect reporting rates to be lowest in predominantly black communities, this does not appear to be borne out by the data. Furthermore, the degree of neighborhood socioeconomic disadvantage is not consistently associated with the likelihood of crime reporting (Baumer, 2002). An association has been observed for simple assaults, but not for robbery or aggravated assault. An extensive review of research in victim crime reporting is given in (Xie and Baumer, 2019).

There are many reasons for why particular incidents may not be reported to police. These include fear of repercussion, victim perceptions that their victimization was ‘trivial’, or might be perceived as such by police, or personal relationships with the offender. Furthermore, there are documented examples of police actively discouraging victims from filing complaints in order to deflate serious crime statistics (Richardson et al., 2019).

### 2.2.3 Predictive policing models

The literature on predictive policing has considered a range of different modeling approaches for spatio-temporal crime forecasting and hot spot selection (Fitzpatrick et al., 2019). To the best of our knowledge, only a small subset of models have been deployed and evaluated in practice.

PredPol, one of the largest vendors of predictive policing software in the US, has been one of few companies to publish modeling details of their hot spot prediction algorithm (Mohler et al., 2011; Mohler, 2014; Mohler et al., 2015). The PredPol algorithm relies on a Self-Exciting Spatio-Temporal Point Process (SEPP) model that uses the location and time of historical incidents to predict the spatio-temporal distribution of future crime within a city. Hot spot predictions for subsequent time steps can be obtained by evaluating the predicted crime distribution on a grid of cells overlaying the city. This model, which has its roots in seismology, separates crime occurrences into “background crime” and “offspring crime” with the rationale that, similar to earthquakes which often trigger close-by aftershock earthquakes, crime tends to form clusters in time and space with burglars returning to the same areas or gang conflicts leading to retaliatory violence (Mohler et al., 2011).

While the SEPP method models both the space and time distribution explicitly, many other common approaches focus of one component at a time. For instance, one straightforward approach is to apply time-series analysis to forecast crime counts in small pre-defined spatial units such as individual segments of streets or grid cells. In a field experiment with the Pittsburgh Bureau of Police, Fitzpatrick et al. (2018) used a within-cell moving average of crime counts in order to predict chronic hot spots and a within-cell multi-layer perceptron on lagged crime count features to predict temporary crime flare-ups. The authors report that the relatively simple moving average model alone was able to capture more crime on average than other models including SEPP models. The Shreveport Police Department in Louisiana conducted experiments with a logistic regression model in 2012 (Hunt et al., 2014). In addition to different lagged crime counts, predictors also included the number of juvenile arrests in the past six month and the presence of residents on probation and parole in each of the 400-by-400-foot grid cells. Other methods focus on the spatial distribution of crime and aggregate the temporal component. Spatial kernel density estimates (KDE's) and risk terrain modelling, which involves risk factors beyond crime rates, are used to help identify chronic hot spots but generally require visual inspection if spatial discretization is to be avoided (Gorr and Lee, 2014; Fitzpatrick et al., 2019). A number of these models including SEPP's (Dulce et al., 2018), KDE's, moving-average type models, and other approaches (Barreras et al., 2016; Dulce Rubio et al., 2018) have previously been applied to historical crime data from Bogotá. Barreras et al. (2016) found, for instance, that KDE models performed the best in their analysis.

In this study, we focus on SEPP models for crime hot spot prediction as they appear to be one of the most widely used and analysed type of model, a trend driven in part by PredPol's popularity and the descriptions of their models publicly available in peer-reviewed literature. For comparison purposes, we also consider a moving average model as analysed in (Fitzpatrick et al., 2018). Both models are based only on the location an time of previous crimes, which makes them particularly accessible to police departments.

## 2.3 Methodology

### 2.3.1 Self-exiting spatio-temporal point processes

Self-Exciting Spatio-Temporal Point Processes (SEPP) are a commonly used class of models for applications in which the rate of events depends on nearby past events, e.g. modeling of earth quakes or the spread of infectious diseases. In the purely temporal case, this class of models is also known as Hawkes processes. We give a short introduction to SEPP, the specifications used in predictive policing and the model used in this study. A more detailed review of SEPP can be found in (Reinhart, 2018).

SEPPs separate events into two types: background events and offspring events. Background events are generally assumed to occur independently across space and time according to a Poisson point process. Each event can then cause offspring events in its vicinity according to a triggering function decaying in space and time. The rate of events at locations $(x, y) \in X \times Y \subseteq \mathbb{R}^2$ and times $t \in [0, T]$ is characterized by the conditional intensity, defined as

$$\lambda(x, y, t|\mathcal{H}_t) = \mu(x, y) + \sum_{\{k:t_k<t\}} g(t - t_k, x - x_k, y - y_k), \tag{2.1}$$

where $\mathcal{H}_t = \{(x_i, y_i, t_i)\}_{i=1}^n$ is the history of events up to time $t$ which we will omit for simplification of notation. The background intensity $\mu(x, y)$ is often assumed to be time-independent while the triggering function $g(t - t_k, x - x_k, y - y_k)$ is generally chosen to be separable in time and space for computational simplicity. For each event $(x_k, y_k, t_k)$, the number of offspring events follows a Poisson distribution with mean

$$m = \int_{X\times Y} \int_T g(t, x, y) \mathrm{d}t \mathrm{d}(x, y).$$

If properly normalized, $g(t - t_k, x - x_k, y - y_k)$ induces the probability distribution of the locations and times of these events. After model fitting, the SEPP can be used to predict the locations and times of future events. Assume we want to predict the number of events $N_{A,t}$ within an area $A \subseteq X \times Y$ at a given time $t = t'$. This prediction can be obtained by computing the integral

$$\widehat{N_{A,t'}} = \int_A \lambda(x, y, t|\mathcal{H}_{t'}, t = t') \mathrm{d}(x, y). \tag{2.2}$$

SEPP models were first applied to crime data for hot spot prediction by Mohler et al. (2011). Initially, the authors suggested non-parametric estimation of $\mu$ and $g$ based on only background or offspring crimes respectively which requires a computationally expensive iterative stochastic declustering procedure. In subsequent work, Mohler (2014) introduced a parametric approach that uses all data to estimate the background intensity with kernel density estimation and assumes a triggering function that is exponential in time and Gaussian in space. The benefit of this parametric approach is that model parameters can be be estimated with a less expensive Expectation-Maximization procedure. In field experiments with the Los Angeles Police Department and the Kent Police Department, United Kingdom, Mohler et al. (2015) forgo a complicated spatial model by fitting a cell-wise constant background intensity and a triggering function only exponential in time.

In this work, we draw on a fully parametric SEPP model that is inspired by the simulations conducted by Mohler et al. (2011). We assume a scaled Gaussian background intensity, defined as

$$\mu(x,y) = \frac{\bar{\mu}}{2\pi(15)^2} \exp\left(-\frac{x^2}{2(15^2)}\right) \exp\left(-\frac{y^2}{2(15^2)}\right), \tag{2.3}$$

where the spatial deviation is chosen purposefully large to ensure support on the whole city map. Our triggering function is similar to the proposed parametric functions and takes the form

$$g(t,x,y) = \theta\omega \exp(-\omega t) \exp\left(-\frac{x^2}{2\sigma_x^2}\right) \exp\left(-\frac{y^2}{2\sigma_y^2}\right). \tag{2.4}$$

Choosing a fully parametric model allows us to analyze a best-case scenario of the bias introduced by differential crime reporting rates as similar models can be used for data simulation and model fitting keeping error introduced by model misspecification at a minimum. In addition, the model choice enables efficient computation of the prediction integrals in Equation 2.2. For crime hot spot prediction, city maps are generally split into small areas by imposing a grid with fixed cell lengths. To predict the number of crimes within a cell at time $t$, integration over the estimated intensity function is necessary which can be computationally expensive depending on the exact model choice. To the best of our knowledge, the model we use is similar to the model employed by PredPol's commercial hot spot prediction software.

### 2.3.2 Expectation-maximization procedure

The parameters of the SEPP model in Equation 2.1-2.4 are estimated using maximum likelihood. As an analytical solution is intractable, Veen and Schoenberg (2008) introduced an Expectation-Maximization (EM) algorithm that maximizes the log-likelihood. Assuming we know the branching structure of the data set $\{(x_i, y_i, t_i\}_{i=1}^n$, i.e. which events were triggered by which previous events and which events come from the background process, we introduce a latent variable $u_i$ which equals $j$ if crime $i$ was triggered by crime

$j$ and $0$ if it was sampled from the background process. Given these latent quantities, the complete-data log-likelihood of the parameter vector $\Theta = (\bar{\mu}, \theta, \omega, \sigma_x, \sigma_y)$ can be written as

$$\begin{aligned} l(\Theta) = &\sum_{i=1}^{n} \mathbb{I}(u_i = 0) \log(\mu(x_i, y_i)) \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}(u_i = j) \log(g(t_i - t_j, x_i - x_j, y_i - y_j)) \\ &- \int_{X \times Y} \int_{T} \lambda(x, y, t) \mathrm{d}t \mathrm{d}(x, y), \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Given a data set of crime events, the EM algorithm provides an iterative procedure of estimating the triggering probabilities $u_i$ and the parameters $\Theta$. In the E step, we estimate the triggering probabilities based on current parameter values as

$$\mathrm{P}(u_i = j) = \begin{cases} \frac{g(t_i - t_j, x_i - x_j, y_i - y_j)}{\lambda(x_i, y_i, t_i)} & \text{if } t_j < t_i, \\ 0 & \text{else.} \end{cases}$$

and $\mathrm{P}(u_i = 0) = \mu(x_i, y_i)/\lambda(x_i, y_i, t_i)$. These latent values can then be plugged into the expected complete-data log-likelihood which gives

$$\begin{aligned} \mathbb{E}[l(\Theta)] = &\sum_{i=1}^{n} \mathrm{P}(u_i = 0) \log(\mu(x_i, y_i)) \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{P}(u_i = j) \log(g(t_i - t_j, x_i - x_j, y_i - y_j)) \\ &- \int_{X \times Y} \int_{T} \lambda(x, y, t) \mathrm{d}t \mathrm{d}(x, y). \end{aligned}$$

In the M step, we maximize the expected log-likelihood with respect to $\Theta$ and return to the E step with the new parameter estimates. This procedure is repeated until the parameter values converge.

### 2.3.3 Bogotá victimization and reporting survey

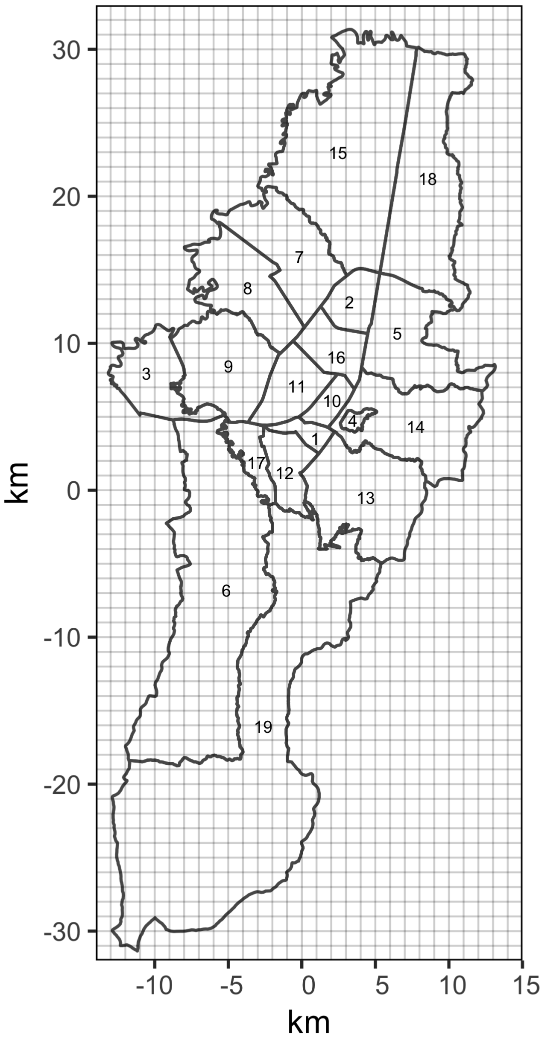
Victimization rates — i.e. the fraction of the population who has been victim of a crime within a given time window—and victim crime reporting rates — i.e the fraction of crime victims who has reported the offense to the police — can generally not be assessed based on only police data but require large-scale surveys. Often, these surveys are not conduced or published with a high-enough spatial resolution to give a sense of differences at a local level. For instance, the US Bureau of Justice Statistics conducts a bi-annual National Crime Victimization Survey with around 95,000 households and publishes rates of victimization and crime reporting on a national level and aggregated by urban, suburban and rural areas (Morgan and Oudekerk, 2019).

In order to study the effect of differential victim crime reporting on predictive policing systems, which are generally limited to a single city, a higher spatial resolution of victimization and crime reporting rates is required. Fine-grained data sets like this are rare and, based on availability, we draw on district-level data from Bogotá, Colombia collected by Bogotá's chamber of commerce, Cámara de Comercio de Bogotá (CCB).

The bi-annual CCB crime perception and victimization survey includes approximately 10,000 randomly selected participants from all socio-economic statuses and all 19 urban districts of Bogotá. Among other questions, participants are asked to indicate whether they have been the victims of a crime in the present calendar year and, if yes, whether they have reported the crime to the police. Results of the surveys are available on the CCB website and are used to inform the definition and adjustment of the city's public policies (de Comercio de Bogotá). Not all of the published reports stratify results by districts. For our experiments, we use victimization and victim crime reporting rates stratified by district based on the survey that covers the first half of 2014 (de Comercio de Bogotá, 2014). Districts, population sizes and rates are depicted in Figure 2.1. Both the crime victimization rates and the victim crime reporting rates vary significantly between different districts with victimization rates between 5 % and 18 % and victim crime reporting rate from 13 % to 33 %. Although the range of both rates can be expected to vary significantly between different cities and countries, this data allows us to analyze the impact of differential crime reporting on predictive policing in a realistic scenario.

### 2.3.4 Synthetic data generation

We simulate location and time of reported and unreported crime incidents in Bogotá districts according to the victimization and victim crime reporting rates displayed in Figure 2.1. In order to minimize possible errors due to model misspecification and instead concentrate on the effect of differential reporting rates, we sample data directly from a high-intensity SEPP and subsample according to each district's victimization



| Id | District | Pop. | Vict. rate | Rep. rate |
|---|---|---|---|---|
| 1 | Antonio Nariño | 109,176 | 15 % | 33 % |
| 2 | Barrios Unidos | 243,465 | 12 % | 22 % |
| 3 | Bosa | 673,077 | 13 % | 26 % |
| 4 | Candelaria | 24,088 | 12 % | 22 % |
| 5 | Chapinero | 139,701 | 9 % | 28 % |
| 6 | Ciudad Bolívar | 707,569 | 8 % | 17 % |
| 7 | Engativá | 887,080 | 11 % | 20 % |
| 8 | Fontibón | 394,648 | 10 % | 19 % |
| 9 | Kennedy | 1,088,443 | 13 % | 28 % |
| 10 | Los Mártires | 99,119 | 17 % | 25 % |
| 11 | Puente Aranda | 258,287 | 14 % | 32 % |
| 12 | Rafael Uribe Uribe | 374,246 | 12 % | 15 % |
| 13 | San Cristóbal | 404,697 | 13 % | 21 % |
| 14 | Santa Fe | 110,048 | 17 % | 17 % |
| 15 | Suba | 1,218,513 | 5 % | 19 % |
| 16 | Teusaquillo | 153,025 | 14 % | 19 % |
| 17 | Tunjuelito | 199,430 | 17 % | 23 % |
| 18 | Usaquén | 501,999 | 18 % | 13 % |
| 19 | Usme | 457,302 | 9 % | 33 % |

Figure 2.1: Bogotá district map with division into 1 km × 1 km grid cells for hot spot prediction and survey-based victimization and victim crime reporting rates. Districts differ notably in size, population numbers, and rates.

rate. The background intensity of the SEPP is a sum over bivariate Gaussian distributions centered at 14 locations spread out evenly on the Bogotá map. Each background crime triggers offspring according to a triggering function that is Gaussian in space and exponential in time coinciding with the model we are fitting (see Equation 2.1-2.4). Since the data will be used to predict hot spots on a fixed grid, we impose a grid of 1 km × 1 km cells on the Bogotá map as depicted in Figure 2.1. District membership of each cell is decided based on its center and each point is attributed to the district of the cell it falls into. We dicretize the time component into daily units and simulate crime data for 2,190 timesteps (6 years) as follows:

1. Sample a set of candidate points $\mathcal{C} = \{(x_i, y_i, t_i)\}_{i=1}^{N}$ from $\lambda$ and discard all points that fall outside of the city bounds or time horizon.

2. For each district $d$ and data within its bounds $\mathcal{C}_d \subseteq \mathcal{C}$, we subsample $n_d \sim \text{Bin}(|\mathcal{C}_d|, p_d)$ of the points to form the true crime data set $\mathcal{D}$, where

$$p_d = \frac{\text{population}(d) \cdot \text{victimization rate}(d) \cdot 12}{|\mathcal{C}_d|}.$$

3. To get a data set of only reported crime, we subsample $n_d \sim \text{Bin}(|\mathcal{D}_d|, q_d)$ crimes for each district $d$ where $\mathcal{D}_d \subseteq \mathcal{D}$ is the set of crimes falling into the given district and

$$q_d = \text{victim crime reporting rate}(d).$$

We implicitly assume that each person is victim of at most one crime which leads to the time scaling factor $2190/(365/2) = 12$ in step 2 as the CCB survey provides rates of victimization for a half-year period. In addition, district population counts are scaled by $1/40$ to speed up the run time of the whole simulation. The described sampling procedure for the true data $\mathcal{D}$ ensures that crime is sampled according to population size and victimization rates but remains distributed according to a thinned SEPP that can be accessed for evaluation of the ground truth conditional intensity. Since $\mathcal{D} \sim p_d\lambda$, the true expected number of crimes in a subarea $A_d$ of district $d$ in time $t = t'$ can be computed as

$$\mathbb{E}[N_{A_d,t'}] = \int_{A_d} p_d \lambda(x, y, t | \mathcal{H}_{t'}, t = t') \mathrm{d}(x, y),$$

where $\mathcal{H}_{t'} = \{(x_i, y_i, t_i) \in \mathcal{C} : t_i < t'\}$.

Figure A.1 depicts a summary of the sampled number of crimes per district, the number of crimes expected according to above integral and the number of crimes as implied by the CCB survey showing that the synthetic data set has the desired rates of victimization for each of the districts.

## 2.4 Results

### 2.4.1 Hot spot prediction procedure

We fit SEPP models (see Equation 2.1-2.4) on the full and reported crime data by discarding the data from the first 500 simulated time steps and training on the subsequent 1,500 days ($\approx$ 4 years) of sampled incidents. Ignoring the first 500 time steps omits the period in which the data generating SEPP is converging to its equilibrium rate and provides a data set that more closely resembles the crime data over fixed time windows we would expect to see in practice. In addition, the time range of approximately 4 years is reasonably close to real crime data sets and falls well within the range of 2-5 years that is suggested by PredPol specifically (PredPol).

The fitted models are used to predict crime intensities on a day-to-day basis for 189 evaluation days where, after each time step, the data for the time step is observed and added to the estimated intensity function for future predictions. On each prediction day, we compute the models' intensity integrals in each of the 1 km × 1 km Bogotá grid cells. These integrals correspond to the absolute predicted crimes per cell and are subsequently used for hot spot selection. Since police are generally only able to patrol small fractions of a city effectively, we select the top 50 cells with highest predicted crime as hot spots which corresponds to approximately 5.7 % of the city's area. Results are aggregated over 50 simulation runs where each simulation samples a new crime data set.

### 2.4.2 Equity between districts

#### 2.4.2.1 Relative number of predicted hot spots

We now discuss the equity of hot spot selection at a district-level. We start by examining how the number of predicted hot spots compares to the number of true hot spots per prediction day in each district. In the case where police are deployed in accordance with the model's predictions, this directly corresponds to the degree of police presence per district relative to a best-case hot spot policing program in which the

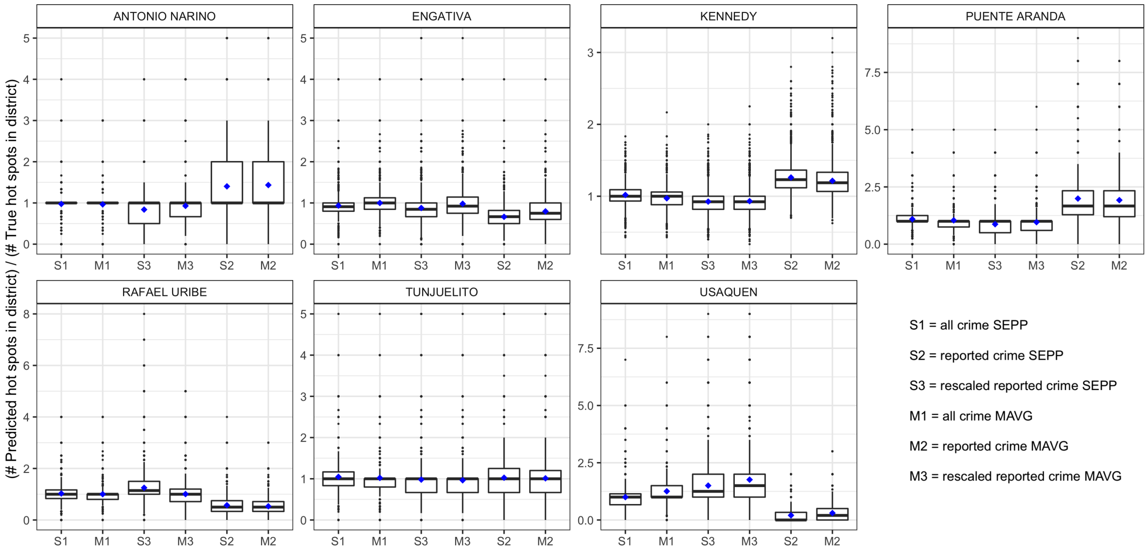


Figure 2.2: Relative number of predicted crime hot spots for a selection of Bogotá districts. Each data point represents a district-specific fraction at a given evaluation time step (189 days) in a given simulation run (50 runs). A total of 50 hot spots are selected at each time step. See Figure A.2 for relative predicted hot spot counts for all districts.

true crime distribution is known.

Figures 2.2 and 2.3 depict the relative hot spot counts for a subset of districts over all evaluation time steps and simulation runs. For Figure 2.2, we set the relative count to 1 for cases in which the district has zero true hot spots and the model correctly predicts zero hot spots and exclude cases with zero true but non-zero predicted hot spots. We see that the SEPP model that was trained on all crime data, i.e. reported and unreported, performs well at selecting the correct number of hot spots uniformly over all districts (S1). This observation is unsurprising given that the fitted model closely resembles the data generating model.

In contrast, the SEPP model that was trained on only reported crime data (S2) is found to have differential performance across districts. Although in some districts, e.g. in Tunjuelito, the relative hot spot counts of the two models appear to be similar, the model with under-reporting on average overestimates the number of hot spots in districts such as Antonio Nariño, Puente Aranda or Kennedy, while underestimating the number of hot spots in districts such as Usaquén, Rafael Uribe Uribe or Engativá. The direction of the introduced error aligns with the victim crime reporting rates of the respective districts as compared to a Bogotá-wide average with fewer of the true hot spots detected in low reporting areas and instead overly many hot spots predicted in high reporting areas. In Usaquén, which with 13 % has the lowest victim crime reporting rate among all districts, only 20.4 % of the number of true hot spots are predicted on average. Meanwhile in Kennedy, which has a comparatively high reporting rate of 28 %, the model on average predicts 126.1 % the number of true hot spots.

Thus far, we have disregarded cases in which none of the true hot spots fall into a given district but the prediction model selects one or more cells. Figure 2.3 gives a summary of the fraction of cases with no true hot spots, further confirming the observed displacement effect of hot spot predictions. In Usaquén, the number of times crime hot spots are predicted when none of the true top 50 crime hot spots lie in the district is over twice as high in the full data SEPP compared to the reported crime SEPP. The same fraction increases more than threefold in the high-reporting district Antonio Nariño, and almost twofold in Puente Aranda. Notably, Figure 2.3 also shows that the displacement effect both impacts districts that almost always have areas with highly concentrated crime and districts that do not. This phenomenon is a function of victimization rates, population sizes and the size of districts.

Finally, average absolute numbers of over- or underpredicted hot spots are displayed in Figure A.4. Although comparison of relative counts ensures that districts of different sizes are evaluated similarly, in some cases we might be interested in the number of grid cells affected by the introduced bias as they roughly relate to the number of impacted individuals. For example, we see that the displacement of predicted hot spots based on differential victim crime reporting rates leads to on average 3.3 too many hot spots predicted in Kennedy while only 0.64 too many cells in Antonio Nariño are selected on average.

Overall, differential reporting rates across districts seem to lead to differentially well-measured aggregate crime levels which distorts the distribution of hot spots. If the police follows the model's recommendations, the consequence would be an unfair allocation of police patrols where areas with low victim crime reporting rates are met with artificially decreased police presence while areas with higher reporting rates are chronically over-policed.

#### 2.4.2.2 Crime threshold for hot spot selection

Calculating relative counts of predicted hot spots gives insights into how much under- or over-policing we can expect per district. A natural way of comparing between districts is to look at the true crime rates required for a cell to be selected as a hot spot. If this threshold is much lower for some districts than for others, the consequence could be more average police presence in these districts despite similar or even

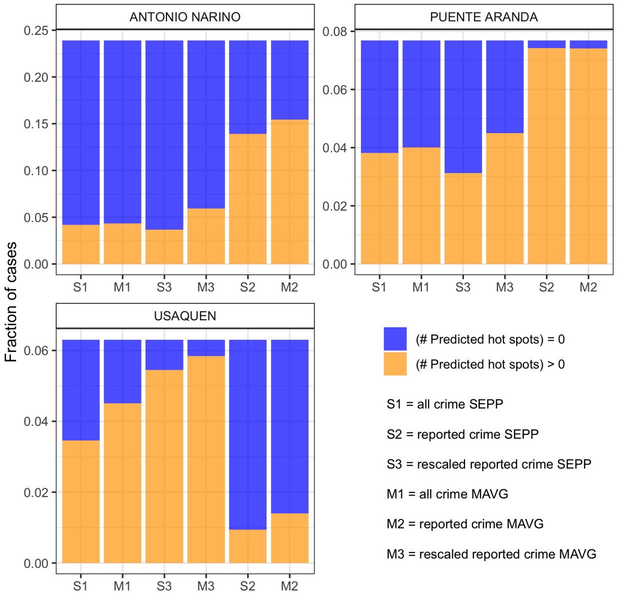


Figure 2.3: Fraction of prediction time steps with no true hot spots in districts. We separate instances into cases with predicted and no predicted hot spots. See Figure A.3 for a version with all districts.

higher crime levels in other areas.

Figure 2.4 shows that the predicted crime rates implied by the reporting data SEPP model present a differentially well-adjusted approximation of true crime rates. Comparing the normalized maps in the Figure, the reported crime SEPP appears to overestimate the relative concentration of crimes in the high-reporting regions Kennedy and Antonio Nariño, and underestimate the relative concentration of crimes in low-reporting districts such as Rafael Uribe Uribe and Usaquén. Moreover, crime rate prediction seems to perform poorly in areas with little true crime. While the ground truth shows clear differences between crime intensities in areas such as Ciudad Bolívar and Usme, the model predictions in these districts appear to be almost indistinguishable.

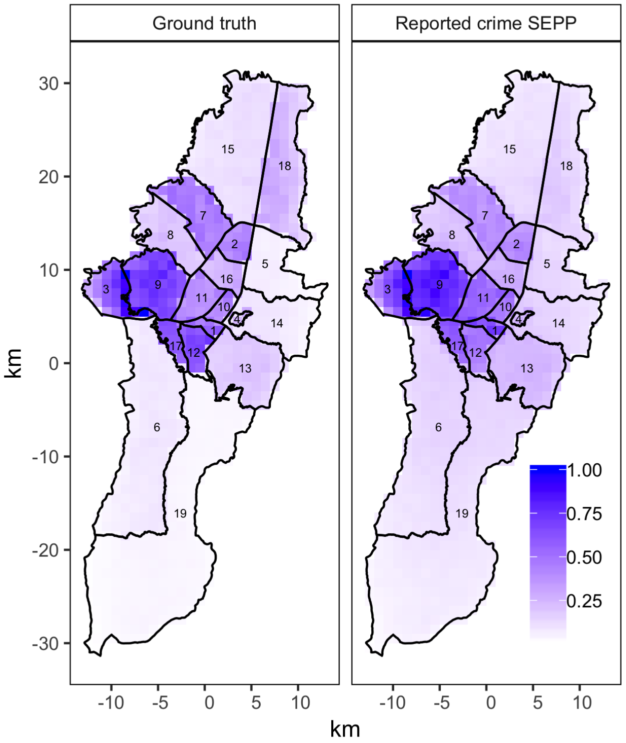
In order to measure equity of model predictions between districts, we consider the minimum true crime rate that leads to a predicted hot spot at each prediction step and summarize the results in Figure 2.5. Since exact crime counts vary over time and this metric omits steps with no predicted hot spots falling into the respective district, the average thresholds have some variability even for full data models. However for districts that are regularly predicted to have hot spots, the full data SEPP model (S1) exhibits very similar hot spot prediction threshold of around 0.5 expected crimes per cell and time step where the low threshold is explained by the population scaling we conducted while simulating Bogotá crime data. In contrast, the model trained on only reported crime data results in varying thresholds even across districts which are regularly predicted to have hot spots. The district-wide average threshold of 0.45 true expected crimes per cell is increased in areas with low crime reporting, e.g. to a rate of 0.73 true crimes on average in Rafael Uribe Uribe and 0.62 in Usaquén. At the higher end of victim crime reporting rates, grid cells in Puente Aranda on average only require a rate of 0.32 true crimes and cells in Kennedy only 0.27 to be selected as a crime hot spots. More concretely, this means that on average the minimum true crime rate that leads to a predicted hot spot in Rafael Uribe Uribe is 2.7 times the minimum crime rate required in Kennedy. In order to rule out the possibility that Kennedy's threshold is artificially high because all of the cells in the district are regularly selected as hot spot, we examine the absolute predicted hot spot counts and find that at no time step more than 72.97 % of Kennedy is selected as hot spot area with a mean of 48.18 %.

These findings imply that crime hot spot prediction in real-world settings with differently sized regions and differential victimization and crime reporting rates can have noticeably biased outcomes that lead to over-policing of some areas of a city while others have higher levels of crime.

### 2.4.3 Scaling by victim crime reporting rates

It is not unusual for police and predictive policing algorithms to leverage data sets beyond registered crime incidents (Shapiro, 2017; Jefferson, 2017; Giménez-Santana et al., 2018). In the case presented here, one could imagine pairing the reported crime data with the survey data to attempt to correct the bias introduced by differential crime under-reporting. Intuitively, this entails rescaling the predicted crime rates according to the reporting rates. Of course, in most cases exact crime reporting rates are unknown to the police. However, as we discuss in this section, even in cases where the reporting rates are known, this rescaling does not necessarily recover the true crime distribution at the finest level.

We explore the rescaling approach as an additional model in our hot spot prediction simulation by taking the integrated intensities in grid cells supplied by the reporting data SEPP and dividing them by the victim crime reporting rate of the respective district. After rescaling, we select the cells with the top 50 highest predictions as hot spots analogous to the other models. The relative predicted hot spot counts of the rescaled model (S3) are displayed along the other models in Figure 2.2. Across the displayed districts, the mean relative number of predicted hot spots is just as close or closer to the number predicted by the full data model (S1) than the reporting data based predictions (S2). This indicates that the rescaling



| Id | District |
|---|---|
| 1 | Antonio Nariño |
| 2 | Barrios Unidos |
| 3 | Bosa |
| 4 | Candelaria |
| 5 | Chapinero |
| 6 | Ciudad Bolívar |
| 7 | Engativá |
| 8 | Fontibón |
| 9 | Kennedy |
| 10 | Los Mártires |
| 11 | Puente Aranda |
| 12 | Rafael Uribe Uribe |
| 13 | San Cristóbal |
| 14 | Santa Fe |
| 15 | Suba |
| 16 | Teusaquillo |
| 17 | Tunjuelito |
| 18 | Usaquén |
| 19 | Usme |

Figure 2.4: Normalized average crime in each cell. The left side depicts the average over true intensity integrals, while the right side uses predictions from the SEPP model trained on only reported crime data. In both cases, we normalize by dividing by the respective maximum average prediction value.

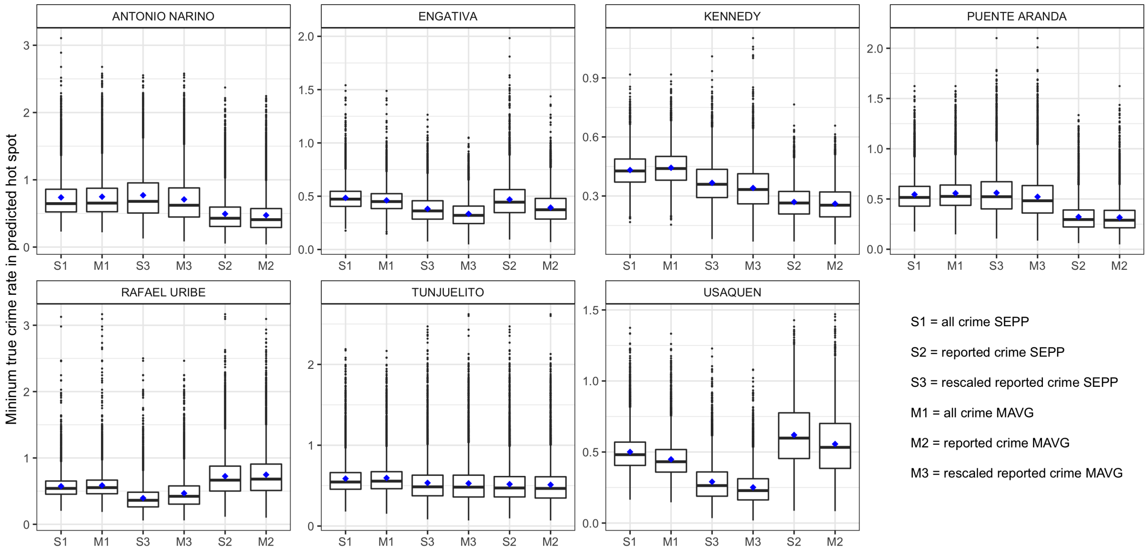


Figure 2.5: True crime thresholds for hot spot selection in a set of Bogotá districts. Each point corresponds to an evaluation time step (189 days) and a simulation run (50 runs). See Figure A.5 for all districts.

strategy successfully reduces outcome disparities. However, this conclusion is called into question when examining the implied minimum true crime rate for hot spot selection shown in Figure 2.5. For example in Usaquén, the rescaled model implies a visibly lower average true crime threshold for hot spot selection than the full data model, and in Engativá the difference between the full data and rescaled models appears to be larger than the difference between the full and reporting data models.

The conflict between the equity measures is observed because the relative predicted hot spot counts are an aggregate metric over all cells and not sensitive to which cells are selected, in contrast to the minimum true crime threshold for hot spot selection. Rescaling of the reporting data SEPP predictions increases predictions in all cells of a district by the same factor without accounting for how much crime was unobserved in each of the cells. As a consequence, the rescaled model selects an approximately correct number of hot spots in many of the districts while the exact cells might not coincide with the true hot spots. In order to recover the cell-wise true crime distribution, a cell-by-cell rate of crime reporting would be required, which presupposes separate representative surveys in hundreds of micro-areas. While incorporating victimization survey information does help to reduce disparities to some extent, it evidently does not suffice in order to fully debias the prediction system.

### 2.4.4 Comparison to a moving average model

In this section we study the behavior of a simple moving average (MAVG) prediction model to assess whether our findings hold more generally outside of the SEPP prediction model setting. MAVG prediction models are fitted analogously to the SEPP models on the full and reported crime data sets. Despite being perhaps the simplest possible prediction model, MAVG's have been found to perform particularly well for detecting long-term hot spots (Fitzpatrick et al., 2018) in real world data.

For our application, we aggregate crime in the same 1 km × 1 km grid cells previously used and fit a within-cell MAVG model to predict the daily crime counts on the same training data sets as before. To obviate the need for tie-breaking that would arise if using simple averaging, we instead employ an exponentially-weighted MAVG model. The same model parameter is estimated for the entire Bogotá grid by searching over a linear scale of bandwidths for the exponential smoothing kernel and selecting the parameter that induces minimal average error with lagged prediction on the training data set. The models are updated on a daily basis by incorporating the crime counts of the previous day.

The performance of the full data MAVG model (M1), the reporting data MAVG model (M2), and the rescaled MAVG model (M3) are depicted in Figure 2.2, 2.3 and 2.5 alongside the corresponding SEPP models. We observe that the MAVG models generally perform similarly to their respective SEPP counterparts. The full MAVG model (M1), for instance, performs on par with the full data SEPP model. Likewise, the reporting data MAVG (M2) induces similar outcome disparities in relative hot spot counts and minimum true intensities as the SEPP trained on victim crime reporting data, and the rescaled MAVG model (M3) struggles to correct the finer resolution bias similarly to the rescaled SEPP model.

At first glance these similarities might be surprising, especially because the true data was simulated from a SEPP. However both the SEPP and the MAVG model follow similar modeling ideas. While the MAVG model forgoes the spatial modeling component by discretizing into grid cells prior to prediction, whereas the SEPP has a continuous underlying intensity that is later integrated over grid cells, both methods model the time between events with an exponential function. In addition, both models make predictions based on a weighted average of previous nearby events and the weights can be fairly similar if we assume that the spatial deviation of the triggering function is small in comparison to the size of the grid cell such that most offspring crimes fall into the same cell as their parent. This assumption is often justified as the criminology literature tends to describe crime hot spots as micro areas of only a few blocks or street segments with high concentration of crime (Fitzpatrick et al., 2019). Indeed, in their randomized controlled field trials, the researchers affiliated to PredPol omit the spatial component of the SEPP altogether and discretize crimes into cells before modeling (Mohler et al., 2015).

## 2.5 Discussion

Our analysis demonstrates how predictive policing systems exclusively trained on victim crime reporting data can lead to spatially biased outcomes due to geographic heterogeneity in crime reporting rates. This in turn can result in over-policing of certain communities while others remain under-served by police.

Our findings are based on synthetic crime data simulated according to district-level victimization and victim crime reporting rates published by Bogotá's chamber of commerce, Cámara de Comercio de Bogotá. We empirically evaluate the equity of predictions across districts of a hot spot prediction algorithm similar to the models used by PredPol. Our findings suggest that districts with low crime reporting rates have fewer of their crime hot spots detected by the algorithm. Conversely, districts with high crime reporting rates are found to have a higher concentration of predicted hot spots than the true crime levels would justify. Moreover, the effective true level of crime required for the model to predict a hot spot is found to vary by more than a factor of two across the districts.

We explore if known victim crime reporting rates can be used to debias hot spot predictions by scaling crime expectations appropriately. The results suggest that this is unsuccessful when reporting rates are known at a district level but hot spots are predicted at a smaller individual cell level since noise introduced by individually thinned crimes is propagated to the rescaled predictions which makes singling out of specific cells in comparison to other cells in the same district difficult.

Prior work has focused on feedback loops and the potential harms of arrest data-based predictive policing systems (Ensign et al., 2018b; Lum and Isaac, 2016). Yet, in practice, predictive policing systems are based on data from victim crime reports (Cheetham, 2019). Our work presents an initial step toward understanding the effect of bias in victim crime reporting data on predictive policing systems. Our analysis demonstrates the importance of considering reporting rate variation when assessing predictive policing systems for potential harms and disparate impacts.

### 2.5.1 Limitations

#### 2.5.1.1 Crime location vs. survey location

Victimization surveys generally provide us with information on crime reporting based on where people live, not based on where crimes occur. On a small scale like a single city, this spatial disparity makes it hard to take survey-based information into account for police allocation. While this limitation of how survey data is collected does not invalidate our findings—in our simulations we treat the reporting rates as reflecting rates of reporting for crimes *occurring* in the given district—it does present a challenge for using such survey data to de-bias predictions in practice. In order for survey data to be useful for this purpose, questions need to ask not only where respondents reside, but also where the victimization(s) occurred.

#### 2.5.1.2 Static reporting rates and potential deterrence effects

Thus far, we do not take the effects of the actual interventions in the form of patrolled hot spots into account. We hypothesize that both victimization rates and victim crime reporting rates can be susceptible to police presence and a model that jointly describes the interplay of crime, reporting rates and police deployment is required for a more complete picture. One component currently omitted is a deterrence effect of policing. Failing to consider such effects could result in the reallocation of police patrols away from neighbourhoods where they are having the intended deterrence effect, precisely because reported crime rates would be lower when police are successful in deterring crime.

### 2.5.2 Implications and generalizability

#### 2.5.2.1 Relationship to socioeconomic advantage

Research on victim crime reporting shows that the decision to report a crime is influenced by the severity of crime (e.g. Greenberg and Ruback, 1992; Goudriaan et al., 2006), victim characteristics (e.g. Slocum, 2017; Hullenaar and Ruback, 2020), and contextual factors such as neighborhood characteristics (e.g. Baumer, 2002; Slocum et al., 2010; Zhang et al., 2007). While we lack information on victim and crime characteristics in the survey data, we are able to speak to a number of socio-technical implications of our results.

Prior research finds links between severe socioeconomic neighborhood disadvantage and lower reporting rates for simple assault incidents (Baumer, 2002). Goudriaan et al. (2006) obtain similar results analysing crime incidents from the Netherlands paired with the Dutch Police Population Monitor survey. Some studies describe a more indirect effect of socioeconomic status on the likelihood of reporting. For example, (Berg et al., 2011) find that victims who are involved in illegal behavior are less likely to report violent acts against them to the police, and this effect is particularly pronounced in disadvantaged neighborhoods. The findings of (Slocum, 2017) suggest that prior police-initiated contact with law enforcement has a negative impact on the reporting of future crime that is amplified for African Americans and poorer individuals. The authors of (Cattaneo and DeLoveh, 2008) study the help-seeking behavior of women who experience intimate partner violence. The study finds that, for the lowest income women, the severity of violence does not predict whether law enforcement is contacted. With increasing income the severity becomes more indicative of victim crime reporting.

In the Bogotá survey data, there appears to be no clear association between reporting rates and socioeconomic advantage at the district level. Ciudad Bolívar, a district with large urban slums that is home to some of the most socioeconomically disadvantaged residents of Bogotá, has a reporting rate of 17%. In line with previous research, this lies well below the average reporting rate across districts of 22.7%. However, Usaquén, the district with the lowest reporting rate of 13%, is also one of the wealthiest districts in Bogotá. We hypothesize that this is in part explained by the spatial clustering of specific crime types. In particular, Usaquén experiences a greater proportion of residential burglary and theft than other districts (de Comercio de Bogotá, 2015; Giménez-Santana et al., 2018). Given that victim crime reporting rates vary based on perceived severity (Xie and Baumer, 2019), this might contribute to a decreased victim crime reporting rate. Additionally, this may also be influenced by intra-district heterogeneity of wealth, as socioeconomically disadvantaged neighborhoods such as El Codito are also located in this district.

There is no simple relationship between socioeconomic level of districts and the geographical disparities induced by the hot spot prediction algorithm. This is in part driven by the observed complexity in the relationship between socioeconomic status and crime reporting at the district level. For example, areas that we project to be over-policed under hot spot policing include the middle class district Antonio Nariño, the lower middle or working class district Puente Aranda and the working and lower class district Kennedy. Areas observed to be under-predicted likewise include districts inhabited by upper, middle, working and lower class residents. Thus our findings *do not* indicate that variation in crime reporting rates systematically disadvantages Bogotá's districts in a manner that falls along socioeconomic lines.

#### 2.5.2.2 Relationship to demographics

Demographic factors such as age (Baumer, 2002; Bosick et al., 2012; Hashima and Finkelhor, 1999; Watkins, 2005), gender (Baumer and Lauritsen, 2010) and race (Xie and Lauritsen, 2012; Avakame et al., 1999) can play a role in victim crime reporting. Desmond et al. (2016) examine the change in Milwaukee's crime reporting rates after public broadcast of police violence against an unarmed black man. They find that, particularly in black neighbourhoods, residents were far less likely to report crime to police following the incident. Ultimately, race and ethnicity are often correlated with socioeconomic status and location, which makes it difficult to identify the direct relationships between demographic variables and victim crime reporting rates (Shapiro, 2017).

Due to data limitations, we are unable to provide an indepth discussion of the relationship between race, ethnicity and predictive disparities in the Bogotá context, as we do not have access to demographic information on the victimization survey participants. A discussion of the Bogotá-specific interplay of race, ethnicity, crime and policing, and how it might generalize to other contexts, thus remains beyond the scope of this work.

#### 2.5.2.3 Generalizability to other jurisdictions

Crime reporting decisions also operate in a macrolevel context encompassing specific cities, local governments or whole nations (Xie and Baumer, 2019). Gutierrez and Kirk (2015) find that, within the US, metropolitan areas with greater proportions of foreign-born or non-US citizens have decreased crime reporting rates, and the results of (Miller and Segal, 2018) suggest that cities with more female police officers have higher rates of victim crime reporting for violent crimes against women. Goudriaan et al. (2004) analyze data from 16 Western industrialized countries and find that differences in crime reporting rates are not entirely explained by crime types, individual and local contexts, but vary with nation-level factors such as the perceived competence of the police at large.

Since our study is exclusively based on survey data from Bogotá, specific findings do not necessarily generalize to other geographies. In particular, while our analysis did not find evidence of a simple relationship between socioeconomic factors at a district level and predictive disparities, results would likely be different in regions—or at resolutions—where socioeconomic factors are more directly associated with reporting rates.

Although the specific spatial distribution and societal implications of the predictive disparities are likely to vary between different jurisdictions, our results suggest that some form of outcome disparity can be expected if victim crime reporting rates have sufficient spatial variation. Such spatial variation is relatively commonplace and can be expected if, for example, the city has some amount of socioeconomic segregation since crime reporting rates vary with neighborhood disadvantage (Xie and Baumer, 2019; Baumer, 2002; Berg et al., 2011).

#### 2.5.2.4 Combining data sources and debiasing

Predictive policing algorithms rely on crime data collected by law enforcement that has repeatedly been found to be flawed, biased or in other ways ‘dirty’ (Richardson et al., 2019). Much of the attention has focused on biases in police-initiated and particularly arrest data, for instance racial bias in drug related arrests or traffic stops in the US (Pierson et al., 2020; Beckett et al., 2006). PredPol acknowledges some of these biases and publicly states that no drug-related offenses, traffic stops or arrest data are used in their prediction system (PredPol, 2017a). Yet there is a lack of transparency as to how the data types that are ‘too biased’ to be included were identified, to what extend other data sources are biased, and which types of biases were considered. To the best of our knowledge, there has been no consideration of reporting biases although their link to socioeconomic, demographic and cultural factors as described in earlier sections is known. Motivated by this problem, we show that, even when predictive policing algorithms only operate on victim crime reporting data, and thereby attenuate the effects of biased police arrest practices, differential victim crime reporting rates can lead to geographically biased prediction outcomes.

In addition to police data, some predictive policing systems incorporate contextual data from other sources (Shapiro, 2017; Giménez-Santana et al., 2018). For example, HunchLab combines public reports of crime and requests of police assistance with data including weather patterns, geographical features, schedules of major events and even moon phases (Shapiro, 2017). In the setting of the hot spot prediction analyzed in this study, one could imagine the proposal to account for the bias introduced by differential reporting rates by scaling model outputs by the survey-based geographically stratified reporting rates. However, our preliminary experimentation suggests that, although in some cases bias can be decreased, a complete mitigation is not possible if the surveyed victim crime reporting rates do not have sufficient spatial resolution. For successful debiasing, we would require close-to-optimal estimates of victim crime reporting rates at the grid cell level, which is impossible to obtain in practice. Ultimately, it is unclear if debiasing victim crime reporting data is any easier than the unsuccessful previous efforts of mitigating bias introduced by arrest data.

# Chapter 3

# Fairness impact of differential feature under-reporting

Based on (Akpinar et al., 2024): Nil-Jana Akpinar, Zachary Lipton, Alexandra Chouldechova. The impact of differential feature under-reporting on algorithmic fairness, working paper / preprint arXiv:2401.08788.

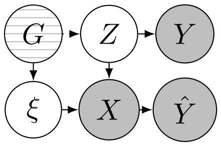
## 3.1 Introduction

Regional and local governments around the world are using their increasingly digitized data systems to develop AI-driven decision-support technologies. The hope is that these tools can help improve decision quality, reduce inefficiencies, eliminate fraud, and improve outcomes for their citizens (Engin and Treleaven, 2019; Levy et al., 2021). Often, these technologies take the form of a predictive risk model. Predictive risk models are prediction models trained on government agencies' administrative data to assess the likelihood that a case will go on to have poor outcomes. Such models have been developed and deployed in criminal justice (Barnes and Hyatt, 2012), child welfare (Vaithianathan et al., 2017), welfare fraud detection (Van Bekkum and Borgesius, 2021), federal tax audits (Houser and Sanders, 2016; Black et al., 2022), homelessness services (Kithulgoda et al., 2022), health care (McCarthy et al., 2015), and many other settings.

Predictive risk models in the public sector have come under criticism over concerns that they are trained on biased data (Mayson, 2019; Richardson et al., 2019; Chouldechova et al., 2018b). In this chapter, we consider one specific form of bias: differential feature under-reporting. We use the term 'differential feature under-reporting' to describe the phenomenon whereby administrative data records are more complete for individuals who have more greatly relied on public services. In the United States, for instance, administrative records often contain medical claims data for those who receive services through public insurance programs (Medicaid and Medicare), but lack information on physical, mental and behavioral healthcare utilization for the privately insured. A lack of *recorded* medical claims information for an individual in this context is typically indistinguishable from instances in which no medical claims have been made.

Differential data availability has been identified as a potential driver of disparities in algorithm-assisted decision-making. For instance, as Eubanks (2018) writes in her critique of the Allegheny Family Screening Tool (AFST) used in screening child maltreatment referrals, ”by relying on data that is only collected on families using public resources, the AFST unfairly targets low-income families for child welfare scrutiny.”

In this work, we provide a technical treatment of this form of data bias and its implications on algorithmic fairness. First, we introduce an expressive yet analytically tractable statistical model of data collection with differential feature under-reporting, and contrast it with other forms of data noise and missingness that have been studied in prior work. We then present novel theoretical results that characterize the impact of differential feature under-reporting on disparities in selection rates across groups. Notably, our results demonstrate that differential feature under-reporting can result in increased *or* reduced disparities. We then describe potential mitigation strategies including adaptations of data imputation methods, and discuss why they generally fail. Instead, we propose a new method based on augmented loss estimation and group-wise imputation that is specifically tailored to the feature under-reporting setting. Lastly, we present empirical results on semi-synthetic and real world data. Our experiments show that the conditions under which disparities decline in the presence of differential data availability rarely arise in practice.



(a) General graph

$G$: High vs. low income group
$Z$: Number of doctor visits in the past year
$\xi$: Publicly insured ($\xi = 1$) or privately insured ($\xi = 0$)
$Y$: Health risk

(b) Illustrative example

Figure 3.1: We study a prediction model on feature vectors with differential under-reporting $X$ where true outcomes $Y$ are a function of the latent unbiased features $Z$. Missingness $\xi$ is influenced by group membership $G$. We consider both cases in which feature distributions vary by group membership and cases with $G \perp Z$. In our setting, missingness indicators $\xi$ are unobserved and group membership $G$ is only used for model evaluation and not as a feature. The graph reflects the dependencies at prediction time.

## 3.2 Preliminaries

We begin by formally describing the problem of differential feature under-reporting, which we illustrate in a directed acyclic graph shown in Figure 3.1. In this setting, an individual's risk prediction, $\hat{Y}$, is formed based on observed administrative data features, $X$, which are a mismeasured version of a ”true” latent feature vector, $Z$. We assume that certain features in $Z$, such as demographic information, are correctly observed in $X$, whereas others, such as use of mental health services, will only be measured correctly for individuals who relied on publicly funded services.

The most challenging aspect of differential feature under-reporting is that administrative data records generally do not distinguish between the absence of information to calculate a feature and a feature being observed as 0. For instance, for indicators and count features — such as indicators of whether a person has recently received inpatient mental health treatment, or a count of the number of episodes in inpatient mental health treatment in the past year — the observed feature will simply show the value 0 for individuals who received those services through privately funded mechanisms. Those who received publicly funded services will not be impacted by the same kind of systematic under-reporting and usually have their indicator/count feature correctly observed. We refer to the phenomenon where in some $Z_{i,j} \neq 0$ appear in the observed data as $X_{i,j} = 0$ as *defaulting* to 0.

Problematically, we generally lack indicators on who is privately or publicly funded and for which services. We know that individuals who have records of reliance on certain publicly-funded services are eligible for those publicly funded services, but we do not know about the rest of the population. This means that when we see a $0$ entry for features $X_j$ subject to differential under-reporting, we do not know whether it's a case where we correctly observed a $Z_j = 0$, or if in actuality $Z_j \neq 0$ and the feature has been mismeasured in the observed data. In the graph, the *unobserved* missingness indicators are denoted by $\xi$. This distinguishes the differential feature under-reporting setting from standard data missingness, wherein the missingness mask, $\xi$, is assumed to be fully observed.

The missing data literature distinguishes three types of mechanisms: (1) Missing Completely At Random (MCAR) where missing values are independent of both observed and unobserved data, (2) Missing At Random (MAR) where missingness depends on observed variables, and (3) Missing Not At Random (MNAR) where missing values are influenced by systematic factors that are not recorded in the data (Rubin, 1976). Under-reporting in administrative data settings does typically not occur completely at random, and instead data is more available for individuals who have relied more greatly on public services which often correlates with demographic attributes like race or income levels. In Figure 3.1, this corresponds to the arrows from $\xi$ to $X$ and $G$ to $\xi$. Since using demographic information in modeling is generally prohibitive, this implies a MNAR setting which is difficult to study in practice. To make the problem more tractable, we assume that feature distributions are the same across groups in parts of this chapter and comment on the more general case whenever possible. This eliminates the arrow from $G$ to $Z$ and effectively renders the setting under-reporting completely at random with unobserved missingness rate.

## 3.3 Background and related work

### 3.3.1 Feature under-reporting and fairness in real-world applications

Data sets with feature under-reporting are ubiquitous across application areas including incomplete administrative data (Eubanks, 2018; Berk et al., 2018), deficient health records (Cismondi et al., 2013; Ahmad et al., 2019), and missing survey data (King et al., 2001; McKnight et al., 2007). That such missingness may lead to bias in predictive models has been pointed out in the context of predictive risk assessment in the public sector (Eubanks, 2018), as well as in the context of health care applications (Rajkomar et al., 2018; Gianfrancesco et al., 2018; Groenwold and Dekkers, 2020). Eubanks (2018) criticizes the Allegheny Family Screening Tool (AFST) used in screening child abuse and neglect referrals by drawing attention to fact that the availability of administrative data is heavily tied to an individual's use of public services.

Table 3.1: Different types of feature mismeasurement and previous work addressing fairness implications. In the data examples, mismeasured features are denoted with $x$ while correctly observed features are denoted by $z$. Columns $g$ encode group membership which may or may not influence mismeasurement of features.

<table>
<tr><th></th><th>Complete data</th><th>Additive noise</th><th>Missing with indicator</th><th>Unobserved missingness</th></tr>
<tr><td><b>Setting and data example</b></td>
<td><table><tr><th>$g$</th><th>$\boldsymbol{z_1}$</th><th>$z_2$</th><th>$y$</th></tr><tr><td>0</td><td><b>10</b></td><td>2</td><td>1</td></tr><tr><td>0</td><td><b>7</b></td><td>1</td><td>0</td></tr><tr><td>1</td><td><b>0</b></td><td>3</td><td>1</td></tr></table><br>• Features fully observed</td>
<td><table><tr><th>$g$</th><th>$\boldsymbol{x_1}$</th><th>$z_2$</th><th>$y$</th></tr><tr><td>0</td><td><b>10.2</b></td><td>2</td><td>1</td></tr><tr><td>0</td><td><b>6.5</b></td><td>1</td><td>0</td></tr><tr><td>1</td><td><b>0.8</b></td><td>3</td><td>1</td></tr></table><br>• Feature values with added random noise $\varepsilon$</td>
<td><table><tr><th>$g$</th><th>$\boldsymbol{x_1}$</th><th>$\boldsymbol{r}$</th><th>$z_2$</th><th>$y$</th></tr><tr><td>0</td><td><b>10</b></td><td><b>1</b></td><td>2</td><td>1</td></tr><tr><td>0</td><td>$\boldsymbol{m}$</td><td><b>0</b></td><td>1</td><td>0</td></tr><tr><td>1</td><td>$\boldsymbol{m}$</td><td><b>0</b></td><td>3</td><td>1</td></tr></table><br>• Some feature values take default value $m$<br>• $r$ indicates which values are observed</td>
<td><table><tr><th>$g$</th><th>$\boldsymbol{x_1}$</th><th>$z_2$</th><th>$y$</th></tr><tr><td>0</td><td><b>10</b></td><td>2</td><td>1</td></tr><tr><td>0</td><td>$\boldsymbol{m}$</td><td>1</td><td>0</td></tr><tr><td>1</td><td>$\boldsymbol{m}$</td><td>3</td><td>1</td></tr></table><br>• Some feature values take default value $m$<br>• No indicators for missingness</td></tr>
<tr><td><b>Previous fairness work</b></td><td>No feature mismeasurement</td><td>Khani and Liang (2020), Phelps (1972), Aigner and Cain (1977), Chen et al. (2018)</td><td>Zhang and Long (2021), Wang and Singh (2021), Jeanselme et al. (2022), Fernando et al. (2021), Fricke (2020), Ahmad et al. (2019)</td><td><b>This work</b>, Eubanks (2018)</td></tr>
</table>

Many of the features used to predict the risk of child abuse or neglect are thus indirect measures of poverty which may lead to inequitable outputs of the risk assessment model. Likewise, electronic health record data is typically missing at different levels for different sub-groups of the population. Missingness in this setting can take the form of defaulting values similarly to the administrative data case. For example, socioeconomically disadvantaged patients may be missing more diagnostic tests compared to others with similar underlying conditions due to limited health care access (Ahmad et al., 2019; Arpey et al., 2017). Reliance on clinical decision support systems trained on data from electronic health records could thus exacerbate already existing health care disparities (Gianfrancesco et al., 2018; Char et al., 2018; Jeanselme et al., 2022).

Similarly, missingness can vary across different domains (e.g. hospitals) which has been studied under the name of domain adaptation under missingness shift (Zhou et al., 2022). The general setup and notation used by Zhou et al. (2022) inspires our problem formulation in this work. While Zhou et al. (2022) consider the problem of different levels of feature missingness between labeled source data and unlabeled target data, we focus on a single domain with varying missingness levels across demographic groups and study the fairness implications of differential feature under-reporting.

### 3.3.2 Feature under-reporting and zero-inflation

Previous literature has studied various notions of under-reporting and potential technical remedies in the context of epidemiology and single-cell RNA sequencing. In epidemiological surveys, social stigma can lead participants to provide false negative responses (e.g. reported maternal smoking $X$ vs. true maternal smoking $Z$) (King et al., 2001; McKnight et al., 2007). Typically, the under-reported features are assumed to be binary with specificity $P(X = 0 \mid Z = 0) = 1$ and sensitivity $P(X = 1 \mid Z = 1) < 1$ (Greenland, 2014; Sechidis et al., 2017). Several lines of work propose methods to estimate the strength of association between the latent feature $Z$ and an observed binary outcome $Y$ by leveraging the observed feature $X$. This includes correction factors for independence tests (Sechidis et al., 2017; Bross, 1954), adjusted mutual information estimators (Sechidis et al., 2017), and corrections for odds-ratio (Chu et al., 2006; Edwards et al., 2013; Dosemeci et al., 1990) and risk-ratio (Rahardja and Young, 2021; Brenner and Loomis, 1994) estimation. These results typically require knowledge of the rate or distribution of missingness which can be obtained from auxiliary validation data. If validation data is not available, domain knowledge is sometimes used to specify possible error distributions in an effort to obtain a range of possible inferences. A few works use a full likelihood approach in which a joint model for $(Z, X, Y)$ is assumed and, under a wide set of assumptions, the unobserved feature $Z$ is marginalized out of the model (Adams et al., 2019). Most of the technical remedies for under-reporting in the context of epidemiological surveys are not applicable to settings with non-binary features and outcomes like the one studied in this work. In addition, the existing methods place the focus on statistical inference rather than prediction. Even in a binary setting with correctly estimated odds ratio, it is unclear how the estimated model can be used at prediction time. This is because, just like at estimation time, only data with under-reporting is available for prediction. Assume we are able to learn a correctly specified model $f(z) = \mathbb{E}[Y \mid Z = z]$ using only mismeasured features $X$ and outcomes $Y$. At prediction time, we observe an example $X = x$ which is potentially impacted by under-reporting. Then $\hat{Y} = f(x)$ is not necessarily the optimal predicted outcome.

A different branch of epidemiological literature studies under-reporting in disease counts in different geographical areas over time. This type of under-reporting can be understood as a form of censoring in which observed counts present a lower bound for true disease counts in an area, that is, instead of defaulting to 0, a feature entry with missing can take any non-negative integer value $X \leq Z$. Most work in this area concentrates on Bayesian modeling with censored likelihood functions (Bailey et al., 2005; de Oliveira et al., 2017) and hierarchical Bayesian models like Poisson-Logistic modeling (Stoner et al., 2019; Gelman et al., 2013)[1]. These methods require a host of parametric and distributional assumptions as well as informative priors or direct access to missingness rates. Assumptions and priors are typically based on in-domain knowledge which is not available in our setting.

In single-cell RNA sequencing, ‘zero-inflation’ or ‘dropout’ refers to the phenomenon by which genes are undetected despite being expressed in a cell due to low levels of RNA. In these cases, the expression level for the gene-cell combination is falsely recorded as zero which needs to be accounted for in further analysis. Corrections are typically incorporated into the dimensionality reduction step (Risso et al., 2018) or conducted via a separate imputation step (Li and Li, 2018; van Dijk et al., 2018; Huang et al., 2018). Both types of methods usually employ Bayesian models and, like the methods for under-reporting in disease counts, require ample in-domain knowledge to justify parametric and distributions assumptions

[1] Similar Poisson-Logistic modeling has also been used in the contexts of econometrics (Winkelmann, 2008), criminology (Moreno and Girón, 1998), and other areas of epidemiology (Greer et al., 2011; Dvorzak and Wagner, 2015)

for gene expression levels.

### 3.3.3 Additive feature noise and fairness

The algorithmic fairness literature has considered the impact of different types of feature mismeasurement as summarized in Table 3.1. A commonly assumed type of mismeasurement is additive feature noise where, instead of a feature $z_1$, we observe a noisy version $x_1 = z_1 + \varepsilon$. The random noise $\varepsilon$ is often assumed to be zero-mean and independent of other variables. Khani and Liang (2020) assume this setting to show that adding the same amount of feature noise to a group-blind model with perfectly predictive features can introduce statistical loss discrepancy when feature distributions vary between groups. This is in line with earlier observations from the statistical discrimination literature. Phelps (1972) and Aigner and Cain (1977) assume a labor market with identically distributed skills between two groups and employers who hire only based on expected skill as measured through some noisy but unbiased test. They observe that different levels of test noise can lead to worse labor market outcomes for one group than the other. The authors remark that the same holds true if the noise levels are the same but the skill distributions vary across groups. In a different line of work, Chen et al. (2018) assume a binary prediction task and propose a decomposition of cost-based fairness metrics into discrepancies based on bias, variance and noise. The authors propose data collection strategies targeted at decreasing the different discrepancy terms and come to the conclusion that overcoming differential noise across protected groups may require measurement of additional features.

The additive feature noise setting is different from the differential feature under-reporting setting studied in this work. Often, additive noise is assumed to have zero mean and small variance as compared to the variance of the impacted feature. This implies that, while some of the information in the feature is diluted, the mean of the feature is not impacted and some of the encoded information remains intact. In contrast, unobserved feature missingness removes all the information from impacted entries and biases the feature mean when the default value does not align with the population mean. Some works (e.g. Khani and Liang, 2020) suggest that feature missingness can be modeled as a special case of additive noise by selecting noise terms with very high variance. While this can successfully simulate settings in which entire feature vectors are removed, the implications for settings in which some features default and some are observed correctly are not clear.

### 3.3.4 Missing data methods and fairness

The intersection of algorithmic fairness and feature mismeasurement in the form of missing observations with indicators has gained increasing attention in recent years. As displayed in Table 3.1, this setting is characterized by feature entries that default to a fixed value (e.g. NaN or a numerical value) while other feature entries are fully observed. Along with the corrupted features $x_1$, we observe a missingness indicator column $r$ with $r = 1$ if $z_1$ in this row has been correctly observed and $r = 0$ if the entry is missing. Missing data of this type is ubiquitous across applications and various methods have been proposed to handle the missing values. In the following, we briefly summarize the literature on the fairness implications of some of these methods.

**Complete case analysis and reweighing** In some cases, it may be desirable to remove observations with missing features from the data entirely and conduct any kind of analyses only on fully observed rows (or columns). This can lead to serious bias in the data when features are not missing at random (Rubin, 1976). To combat some of the problems, various reweighing procedures have been proposed. Zhang and Long (2021) suggest learning only from complete observations while employing an importance sampling procedure with weights representing the normalized inverse of propensity scores. The authors derive theoretical bounds on accuracy discrepancy in terms of the group-wise total variation distances between the weighted distribution of complete examples and the complete data distribution. Experiments by Wang and Singh (2021) suggest that reweighing and resampling methods in the context of categorical data can lead to considerable fairness improvements over learning with missing data directly.

**Imputation** Imputation strategies fill in missing values using methods that vary significantly in complexity (Gondara and Wang, 2018; Li et al., 2019; Van Buuren and Oudshoorn, 1999; LeMorvan et al., 2021). Several recent works empirically examine fairness implications of missing feature imputation in different settings. Jeanselme et al. (2022) compare different imputation strategies under clinical presence and find that there is no imputation strategy that reliably outperforms other imputation methods in terms of fairness. Fernando et al. (2021) and Fricke (2020) compare feature imputation to omission of observations with missing features and find that rows with missing values can contribute to fairer outcomes via observed columns and should therefore not be discarded for training. Ahmad et al. (2019) describe how imputation has the potential to lead to harmful outcomes when used in high-stakes settings like recidivism prediction.

The main difference between the studied settings and the differential feature under-reporting setting considered in this work is that no indicators for missingness are observed in our setting. This makes application of the discussed methods difficult. Despite these difficulties, we experiment with both omitting mismeasured features columns and some adapted imputation methods in Section 3.8.

## 3.4 Problem setup

### 3.4.1 Technical setting

We study the effect of unobserved feature missingness on algorithmic fairness through the lens of the regression setting displayed in Figure 3.1. The general setup and notation are inspired by Zhou et al. (2022) who, instead of studying fairness across demographic groups with varying levels of missingness, focus on domain adaptation under missingness shift. Assume we have latent feature vectors $z \in \mathbb{R}^d$ and group information $g \in \{0, 1\}$. We assume we are in a noiseless regression setting where the outcome $y$ is a linear function of $z$, i.e. $y = \alpha + \beta^T z$ for fixed parameters $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$. We assume that $\beta_i \neq 0$ for $i \in [1 : d]$. Instead of the true features $z$, we observe a mismeasured vector $x$ in which entries default to 0 with group-dependent probabilities. We set

$$x = z \odot \xi^g,$$

where $\odot$ denotes element-wise multiplication, $\xi^g \sim \text{Bern}(m^g)$, and $m^0, m^1 \in (0, 1]^d$ are the rates at which features are observed in the two groups. More formally, we have a group random variable $G \sim \text{Bern}(r)$ where $r$ is the share of the population in group 1, along with a random feature vector $Z$ whose distribution may or may not depend on the value of $G$. Then, the random vector of mismeasured features $X$ can be written as $X = Z \odot \xi$ where $\xi = G * \xi^1 + (1 - G)\xi^0$. This setting allows for two different dependence structures. In the most general version, feature distributions can vary across groups which implies that the missingness random vector $\xi$ and the feature random vector $Z$ are dependent. Since this setting can be difficult to study analytically, we focus on the simplified setting where $G \perp Z$ in some parts of the analysis.

### 3.4.2 Thresholded prediction

The goal of this work is to understand the effect of unobserved missingness on the fairness of a downstream prediction model. The bias differential under-reporting introduces in this context is two-fold: (1) Under-reporting in training data influences the estimation of the prediction model (estimation step), and (2) input data with under-reporting leads to biased predictions at test and deployment time (prediction step). Both training and test data are drawn from the same distribution with the same missingness mechanism. We note that, in general, it is not sufficient to recover the true model parameters $\alpha, \beta$ in this setting as only biased features are available at prediction time. In fact, our experiments in Section 3.9 demonstrate that using the true model parameters for prediction with biased features can lead to worse fairness outcomes than using a model estimated with biased data directly.

We assume a thresholded linear prediction setting reminiscent of predictive risk modeling in the public sector. A linear prediction function $f$ is fit on the pairs $(X, Y)$ to produce predictions $\hat{Y} = f(X) = f(Z \odot \xi^G)$. This only relies on observed features with under-reporting, and does not use the protected group $G$ as a modeling feature with is usually prohibitive. We then consider the group-wise shares of predictions $\hat{Y}$ that lie above a given threshold $\tilde{y}$: $\mathbb{P}(\hat{Y} \geq \tilde{y} \mid G = g)$. We will refer to these group-wise shares as *selection rates* at threshold $\tilde{y}$. Without loss of generality, this setup only considers predictive risk assessment settings in which the *highest* risk individuals are selected (e.g. child welfare screenings, fraud detection, federal tax audits). However, it is straightforward to reverse the analysis and narrative for risk assessment instruments in which the *low* risk leads to selection (e.g. bail decisions in criminal risk assessment). In addition, we will implicitly assume that being selected is *undesirable* and the group that is over-selected is *disadvantaged* if not specified otherwise. This simplifying assumption is made to avoid any confusion in the interpretation of results. Yet, the results presented in this work are sufficiently general to also draw conclusions for the case in which selection is desirable.

The unobserved differential feature missingness setting differs from previously studied feature missingness scenarios primarily in the fact that feature entries are not clearly marked as missing, and instead default to a value that is generally indistinguishable from the correctly observed entries. Since many of the features in administrative data are counts, e.g. the number emergency room visits in the last year or the number of prior offenses, a default value of 0 is a natural choice in this setting. Consider the illustrative example given in Figure 3.1. Here, a government agency is modeling a health-related risk score using features including the number of doctor visits in the past year. Information on health care utilization is routinely available for individuals supported by Medicaid and Medicare, but often missing for privately insured individuals. When comparing a group of individuals with high incomes ($G = 1$) to a group with lower incomes ($G = 0$), this may lead to different levels of feature missingess since the insurance type is tied to income levels.

### 3.4.3 Excess selection rates

Assume a setting in which each observation corresponds to an individual, and we want to select the top $C \in [0, 1]$ share of the population by thresholding outputs $\hat{Y}$ of a prediction model. We argue about the effect of unobserved feature missingness on the equity of the selection procedure by comparing selection rates across groups: that is, comparing the proportion of individuals in each group whose predicted score exceeds the given threshold. Given the cumulative distribution function of predictions $F_{\hat{Y}}$, the percentile threshold $C$ implies an absolute threshold $\tilde{y} = F_{\hat{Y}}^{-1}(1 - C)$ such that the selection rate for group $g$ can be written as $P(\hat{Y} \geq \tilde{y} \mid G = g)$.

In order to isolate the effect of feature missingness, a comparison of selection rates under mismeasurement needs to account for a potential difference in selection rates when features are correctly observed. This difference occurs when the distributions of the true latent features $Z$ vary across groups, which is common in practice. We compare the predictions of a model trained on mismeasured features $X$ as denoted by $\hat{Y}_X$, and the predictions of a model trained on correctly measured features $Z$, denoted by $\hat{Y}_Z$. Note that, since the distributions of $\hat{Y}_Z$ and $\hat{Y}_X$ generally differ, the implied threshold on the model with correctly measured features $y' = F_{\hat{Y}_Z}^{-1}(1-C)$ is generally not the same as the threshold using mismeasured features $\tilde{y} = F_{\hat{Y}_X}^{-1}(1 - C)$.

With this notation in hand, we are able to define an equity-related metric that allows us to assess the impact of differential feature under-reporting on disparities in selection rates. Note that, in principle, we could directly consider a ”difference in differences”: the difference in selection rates between groups $g = 0, 1$ when selection occurs according to the model $\hat{Y}_X$ versus the unbiased predictions $\hat{Y}_Z$. However, since we select a fixed share of the population $C$, an increase of the selection rate of group $g$ when moving from $\hat{Y}_Z$ to $\hat{Y}_X$ already implies a decrease for group $1-g$. We can therefore measure the effect of unobserved feature missingness on selection rates as follows.

**Definition 1** (Excess selection rate due to missingness)**.** *The excess selection rate for group $g \in \{0,1\}$ at overall selection rate $C \in [0,1]$*

$$\Delta(g,C) := P(\hat{Y}_X \geq \tilde{y} \mid G=g) - P(\hat{Y}_Z \geq y' \mid G=g),$$

*is the difference in selection rates when ranking according to a model trained on mismeasured features $X$ compared to a model trained on the correct features $Z$. We say that group $g$ is over-selected at level $C$ if $\Delta(g,C) > 0$. If $\Delta(g,C) < 0$, we say that $g$ is under-selected.*

Note that over-selection of group $g$ implies under-selection of group $1-g$ and vice versa.

It is generally difficult to argue about the excess selection rate $\Delta(g,C)$ analytically. Even in a simple setting with group-dependent Gaussian features $Z \mid G \sim \mathcal{N}(\mu_G, \Sigma_G)$, there is no closed-form expression for the quantile $y' = F^{-1}_{\hat{Y}_Z}(1-C)$ and determining the sign of $\Delta(g,C)$ requires analysis of a difference in cdfs which is often intractable. Instead, we simplify the setting and assume that $Z$ follows the same distribution across groups. In this case, the selection rates on the true outcome $Y$ are the same in both groups at every threshold, and we can simplify Definition 1 as follows.

**Definition 2** (Excess selection rate due to missingness, independent case)**.** *If $Z \perp G$, we say that group $g \in \{0,1\}$ is over-selected at threshold $C \in [0,1]$ if*

$$P(\hat{Y}_X \geq \tilde{y} \mid G=g) > P(\hat{Y}_X \geq \tilde{y} \mid G=1-g),$$

*and under-selected if the inequality is reversed.*

While the majority of our theoretical derivations assume the special case of $Z \perp G$, the empirical portion of this work explores the impact of unobserved feature missingness on selection rates in the more general setting.

## 3.5 Differential feature under-reporting in linear regression

In this section, we examine the bias that differential feature under-reporting introduces into parameter estimates in linear regression. We consider a setting in which true outcomes are a linear function of the latent feature vectors, i.e. $Y = \alpha + \beta^T Z$, which implies that a linear model with access to the true $Z$ recovers the true outcomes $Y$. To simplify notation, we drop the subscripts and write $Y = \hat{Y}_Z$ and $\hat{Y} = \hat{Y}_X$ for the remainder of the paper. We discuss how introduction of feature missingness leads to an attenuation effect in the respective regression parameter and analyze how the model shifts weight to other features when encountering differential feature under-reporting.

### 3.5.1 Estimates and attentuation bias

Feature mismeasurement in the form of unobserved missingness leads to inconsistent parameter estimates in linear regression. When fitting a linear model on $(X, Y)$, the least squares estimates become

$$\begin{aligned} \hat{\beta} &= \Sigma_X^{-1}\Sigma_{XZ}\beta, \\ \hat{\alpha} &= \alpha + \mathbb{E}\left[Z\right]^T \beta - \mathbb{E}\left[X\right]^T \hat{\beta} = \alpha + (\mathbb{E}\left[Z\right]^T - \mathbb{E}\left[X\right]^T \Sigma_X^{-1}\Sigma_{XZ})\beta, \end{aligned} \tag{3.1}$$

where $\Sigma_{XZ}$ denotes the covariance matrix between vectors $X$ and $Z$ and we write $\Sigma_X$ for $\Sigma_{XX}$.

At first glace, this solution resembles the regression estimates in the more commonly studied additive feature noise case. Assuming $X' = Z + U$ where $U$ is independent zero-mean feature noise, the estimate for $\beta$ is $\hat{\beta} = \Sigma_{X'}^{-1}\Sigma_{X'Z}\beta = (\Sigma_Z + \Sigma_U)^{-1}\Sigma_Z\beta$. The factor $\lambda = (\Sigma_Z + \Sigma_U)^{-1}\Sigma_Z$ is commonly interpreted as a noise-to-signal ratio and, if $Z$ is one-dimensional, we know that $\mid \hat{\beta} \mid = \lambda \mid \beta \mid < \mid \beta \mid$ which is generally refered to as attentuation bias (e.g. Hausman, 2001; Fuller, 1987).

In the feature missingness case, the covariance $\Sigma_X$ is not easily separated into terms depending on only the feature or only the mismeasurement. However, in the special case of one-dimensional features and $Z \perp \xi$, we can still show that the parameter $\hat{\beta}$ is biased towards zero.

**Lemma 3** (Attentuation bias)**.** *Assume the feature $Z$ is one-dimensional and has the same distribution across groups ($G \perp Z$). Then the least squares regression of $Y$ on the mismeasured feature $X$ yields an estimated slope $\hat{\beta}$ with $\mid \hat{\beta} \mid \leq \mid \beta \mid$.*

Note that this result does not hold in general if the unobserved missingness, $\xi$, is correlated with $Z$.

### 3.5.2 The $d$-dimensional case

Lemma 3 gives an important insight into the effect of unobserved missingness on parameter estimation in a single feature setting. However, real-world prediction settings typically come with an array of different features that tend to be correlated in some manner. In an effort to understand how the feature correlation structure contributes to the problem of unobserved missingness, we consider the following prediction setting.

Assume the feature vector $Z$ is $d$-dimensional, and missingess only occurs in the first feature. This means that the mismeasured vector $X$ coincides with $Z$ in all but the first entry which is computed as $X_1 = Z_1\xi_1$ where $\xi_1 = G\xi_1^1 + (1-G)\xi_1^0$ and $\xi_1^1 \sim \text{Bern}(m_1^0)$, $\xi_1^0 \sim \text{Bern}(m_1^1)$. Formally, we set $m_i^0 = m_i^1 = 1$ for all $i \in [2:d]$ to denote the setting wherein features $2:d$ are fully observed. We further assume that features $Z_2, \ldots, Z_d$ are uncorrelated and the feature dependence structure is characterized entirely by the correlations between the observed features and the mismeasured feature $\rho(Z_i, Z_1)$ for $i \in [2:d]$. In practical settings, this may be achieved via feature orthogonalization. We explicitly exclude cases in which $Z_1$ is a perfect linear combination of other features to avoid problems of multicollinearity and assume $\mathbb{V}\left[Z_k\right] > 0$ for all $k \in [2:d]$.

**Proposition 4.** *In the described setting, the parameter estimates from Equation 3.1 take the form*

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 \frac{1}{1-R^2}\sqrt{\frac{\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[X_1\right]}}\left(\rho(X_1, Z_1) - \sum_{i=2}^{d}\rho(X_1, Z_i)\rho(Z_1, Z_i)\right), \\
\hat{\beta}_k &= \beta_1\sqrt{\frac{\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_k\right]}}\left(\rho(Z_k, Z_1) - \frac{1}{1-R^2}\rho(X_1, Z_k)\left(\rho(X_1, Z_1) - \sum_{i=2}^{d}\rho(X_1, Z_i)\rho(Z_1, Z_i)\right)\right) + \beta_k
\end{aligned} \tag{3.2}$$

*for* $k \in [2:d]$. *Here,* $R^2 = \sum_{i=2}^{d}\rho(X_1, Z_i)^2 \in [0, 1)$.

The value $R^2$ is the squared coefficient of multiple correlation between $Z_1\xi_1$ and $Z_{[2:d]} = [Z_2, \ldots, Z_d]$ which can be interpreted as the fraction of variance in $Z_1\xi_1$ that can be explained by the independent variables $Z_{[2:d]}$. If all features are observed, the factor $\rho(X_1, Z_1) - \sum_{i=2}^{d}\rho(X_1, Z_i)\rho(Z_1, Z_i)$ collapses to $(1 - R^2)$, and the estimates are unbiased. If some feature values are missing, the bias introduced into the parameter estimates depends on the strength of correlations between features, as well as how this correlation changes with the mismeasurement of $Z_1$. The bias in Equation 3.2 can be conceptualized as a generalization of omitted variable bias (Angrist and Pischke, 2008) which is further explored in Appendix B.2.

We now turn towards the special case in which feature distributions are independent of group membership which implies that whether a value is missing or observed is independent of its latent value. This resembles the assumptions made in previous work on the impact of additive feature noise in fairness (e.g. Khani and Liang, 2020; Phelps, 1972; Aigner and Cain, 1977), and allows us to gain analytical insights that would otherwise remain intractable. First, we examine the behavior of the parameter estimate for the feature with unobserved missingness, i.e. $\hat{\beta}_1$.

**Proposition 5** (Properties of $\hat{\beta}_1$)**.** *If $Z \perp G$, the parameter estimate $\hat{\beta}_1$ has the following properties.*

1. ***Sign invariance:*** *$\hat{\beta}_1$ has the same sign as $\beta_1$.*
2. ***Attentuation bias:*** $\mid \hat{\beta}_1 \mid \leq \mid \beta_1 \mid$.
3. ***Attentuation bias increasing with missingness:*** *If missingness $1 - m_1^g$ is increasing for one (or both) groups $g \in \{0, 1\}$, ceteris paribus, the magnitude of the parameter estimate $\mid \hat{\beta}_1 \mid$ is decreasing.*

This finding shows that missingness in a feature still leads to attenuation bias in the respective parameter estimate, even when other correlated and fully observed features are available. This attenuation bias gets more pronounced with more missingness. Next, we study the properties of the estimates for the fully observed features $\hat{\beta}_k$ for $k \in [2 : k]$.

**Proposition 6** (Properties of $\hat{\beta}_k$)**.** *If $Z \perp G$, the parameter estimates $\hat{\beta}_k$ for $k \in [2 : d]$ have the following properties.*

1. ***Correlation bias:*** *If $\hat{\beta}_k \neq \beta_k$, then $\rho(Z_1, Z_k) > 0$.*
2. ***Shifting weight:*** *If missingness $1 - m_1^g$ is is increasing for one (or both) groups $g \in \{0, 1\}$, ceteris paribus, $\hat{\beta}_k$ is*
   - *increasing if sign $(\beta_1 Cov [Z_1, Z_k]) = +1$, and*
   - *decreasing if sign $(\beta_1 Cov [Z_1, Z_k]) = -1$.*

In line with general intuition, missingness in $Z_1$ has no effect on the parameter estimate $\hat{\beta}_k$ if features $Z_k$ and $Z_1$ are uncorrelated. If the features are correlated, the direction of the effect of missingness on the parameter estimate depends on the signs of $\beta_1$ and Cov $[Z_1, Z_k]$. Note that this is independent of the value and sign of $\beta_k$.

In summary, Proposition 5 and 6 tell a compelling story about the effect of unobserved missingness in the studied setting. As more missingness is injected, the regression model places less and less weight on the mismeasured feature and instead shifts weight to fully observed features with non-zero correlation. This can lead to increasing or decreasing parameter estimates depending on the exact setting. Analytically, there can be cases in which 'shifting weight' means that the absolute value of a parameter estimate $\mid \hat{\beta}_k \mid$ is decreasing. For example, consider a setting in which both $Z_1$ and $Z_k$ have positive true parameters but negative correlation, i.e. $\beta_1, \beta_k > 0$ and Cov $[Z_1, Z_k] < 0$. In practice, this could occur when there are several mutually exclusive paths to the same outcome. For example, consider prediction of general health risk scores with features including both the number of pediatrician visits in the last year and the number of internist visits. Presumably these features are negatively correlated because they are only relevant for two mutually exclusive parts of the population, i.e. children and adults, but in both columns larger values can be indicative of a high general health risk.

## 3.6 Impact on selection rate disparity

### 3.6.1 Selection rate disparity in Gaussian setting

We set out to study the effect of differential feature under-reporting on selection rate disparities in linear regression. While the focus of Section 3.5 was to understand how feature missingness impacts parameter estimates, we now turn towards examining the circumstances under which a group is over-selected or under-selected in the thresholded prediction setting. For the remainder of this section, we assume that the distribution of features $Z$ is independent of the group membership $G$. Following Definition 2, group $g \in \{0, 1\}$ is over-selected at percentile threshold $C$ due to unobserved feature missingness if and only if

$$P\left(\hat{Y}_X \geq F_{\hat{Y}}^{-1}(1-C) \mid G = g\right) > P\left(\hat{Y}_X \geq F_{\hat{Y}}^{-1}(1-C) \mid G = 1-g\right).$$

Similar to our previous discussion, we now assume a $d$-dimensional feature setting in which only the first feature is subject to hidden missingness while all other features are fully observed. We further assume that features are jointly Gaussian, i.e. $Z \sim \mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$

is positive definite. This has the benefit that predictions $\hat{Y} = \hat{Y}_X$ follow a Gaussian mixture distribution which allows us to directly analyze group selection rates. If missingness rates are the same across groups, i.e. $m_1^0 = m_1^1$, there is no selection rate disparity as both groups have the same distributions of features and predictions. If the missingness rates vary between groups, we observe the following.

**Proposition 7.** *Define the threshold turning point $T$ as*

$$T = \hat{\alpha} + \hat{\beta}_{[2:d]}^T \mu_{[2:d]} + \frac{sd\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]}\right)}{sd\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]}\right) - sd\left(\hat{\beta}^T Z\right)} \hat{\beta}_1 \mu_1.$$

*Then, for a high threshold $\tilde{y}$ with $\tilde{y} > T$, the group with more missingness is*

- ***Case 1:*** *Over-selected if* $\mathbb{V}\left[\hat{\alpha} + \hat{\beta}_{[2:d]} Z_{[2:d]}\right] > \mathbb{V}\left[\hat{\alpha} + \hat{\beta} Z\right]$, *or*
- ***Case 2:*** *Under-selected if* $\mathbb{V}\left[\hat{\alpha} + \hat{\beta}_{[2:d]} Z_{[2:d]}\right] < \mathbb{V}\left[\hat{\alpha} + \hat{\beta} Z\right]$.

*For low thresholds $\tilde{y} < T$, the cases are reversed.*

The proposition shows that the question of over-selection primarily depends on the variance in predictions. When cutting off at a high threshold, the group with more missingness is over-selected if the variance in predictions for examples with feature missingness exceeds the prediction variance for fully observed examples (Case 1). It is under-selected if the variance of predictions is larger for the examples with fully observed features (Case 2).

The intuitive narrative is that information deficiency in a group leads to under-selection of that group in thresholded prediction settings as the group's risk distribution concentrates more closely around its mean. In our setting, we find that analytically outcome disparities can go into either direction and do not necessarily lead to under-selection of the group with information deficiency. While our findings suggest that this is mostly a question of variance in predictions, this is likely only part of the story in more general cases with group-dependent feature distributions. We explore how unobserved feature missingness causes over-selection and under-selection of groups in more general empirical settings in Section 3.8.

In practical applications, thresholds are usually set such that only a small portion of predictions exceeds the threshold. For example, we can only decide to flag a small portion of calls as high-risk in the child welfare setting. In particular, realistic thresholds are generally well above the average $\hat{Y}$. On a high level, the turning point $T$ in Proposition 7 represent an adjusted mean predicted value where the influence of the feature with missingness is weighed depending on a ratio determined by prediction variances with and without the feature.

### 3.6.2 Combining parameter estimation and prediction steps

Unobserved feature missingness introduces bias into predictions in two main steps. First, bias is introduced via inconsistent parameter estimates (Section 3.5.1 and 3.5.2). Second, missing features in the prediction step have the potential of introducing additional bias (Section 3.6.1). In the following, we combine the findings of our previous analyses in order to study under what conditions on the true model and features, groups with hidden feature missingenss are over-selected or under-selected. For this, we assume a $d$-dimensional feature setting in which only the first feature $Z_1$ is impacted by missingness. Features are jointly Gaussian, i.e. $Z \sim \mathcal{N}(\mu, \Sigma)$, and we further assume that $Z_2, \ldots, Z_d$ are uncorrelated.

**Corollary 8.** *Given the first and second moments of $Z_1$, the expected share of observed values $\mathbb{E}[\xi_1]$, and the fraction of variance in $Z_1$ that is explained by the remaining features $S^2 = \sum_{i=2}^{d} \rho(Z_1, Z_i)^2$, there exists a positive constant $c = c(\mathbb{E}[Z_1], \mathbb{V}[Z_1], \mathbb{E}[\xi_1], S^2)$ such that, at high thresholds, the group with more missingness is*

- ***Case 1:*** *Over-selected if*
$$\frac{1}{\beta_1}\sum_{j=2}^{d}\beta_j \mathit{Cov}[Z_1, Z_j] < -c,$$
- ***Case 2:*** *Under-selected if*
$$\frac{1}{\beta_1}\sum_{j=2}^{d}\beta_j \mathit{Cov}[Z_1, Z_j] > -c.$$

Thresholds are considered high if they exceed the turning point defined in Proposition 7. The corollary shows that, in the Gaussian case, the question of which group is over-selected due to unobserved feature missingness depends on the signs and magnitudes of the true parameters $\beta$ and the covariances between the features. If

$$\operatorname{sign}\left(\frac{1}{\beta_1}\sum_{j=2}^{d}\beta_j \mathrm{Cov}[Z_1, Z_j]\right) = 1,$$

e.g. if all true parameters and covariances are non-negative, the group with more missingness will always be under-selected at high thresholds (Case 2). If the sign is negative, the group with more missingness is over-selected only if the covariance-weighted sum of true parameters divided by the true $\beta_1$ is sufficiently large in absolute value. Otherwise, unobserved feature missingness still leads to under-selection.

## 3.7 Solution approaches

Sections 3.5 and 3.6 show how ignoring the problem of differential feature under-reporting can lead to models with unfair selection rates across groups. In the following, we explore approaches for mitigating this unfairness including conventional omission and imputation strategies for missing data (Section 3.7.1). We then proceed by proposing a set of methods tailored to the differential under-reporting setting specifically (Section 3.7.2).

### 3.7.1 Standard approaches for handling missing data

Complete case analysis and imputation are two of the most common types of methods for handling data with missing feature entries. Both typically assume that missing entries are clearly marked as missing. Although this assumption is not met in the differential feature under-reporting setting, we discuss the applicability of standard approaches for handling missing data in the following. As before, we assume feature vectors to be $d$-dimensional with missingness in the first feature $Z_1$ which is observed as $X_1 = Z_1\xi_1$.

**Omission of feature** A simple idea to mitigate the bias introduced by feature under-reporting is to discard the mismeasured feature vector $X_1$ entirely. This does not require missingness indicators and can be a viable option if we know the proportion of missing entries to be high, if the feature has low relevance for the prediction target, or if other available features are highly correlated to the mismeasured feature. However in general, omission of entire feature columns can lower model performance significantly and is thus typically avoided. While removing an under-reported feature combats mismeasurement in the feature, the overall impact on outcome fairness is not obvious especially if the feature has high relevance for the prediction target. For example in the linear case with $Y = \alpha + \beta^T Z$, removing $X_1$ leads to omitted variable bias in parameter estimates $\hat{\alpha}, \hat{\beta}_{[2:d]}$ which is explored in Appendix B.2. The direction and magnitude of the parameter estimation and prediction bias then depends on the correlation matrix of features.

**Multiple imputation** Imputing missing values using available data avoids discarding valuable information. Multiple imputation methods draw $m$ plausible values from a distribution, conduct the desired analysis with all $m$ completed data sets, and then pool results to draw conclusions. Since this can be a computationally expensive procedure, the number of imputation runs $m$ is usually kept small (e.g., between 5 and 20). As opposed to single imputation values, multiple imputation retains variability and decreases potential bias introduced through imputation. Since no missingness indicators are observed in our setting, we experiment with imputation of all 0-entries in $X_1$. Crucially, this also includes entries that have been correctly observed as 0. Imputed values are modeled using observed features $Z_2, \dots, Z_d$ but not the target $Y$ since the imputation step has to be conducted at prediction time and $Y$ is only available for training data. In each imputation run, we estimate the posterior $P(Z_1 \mid Z_2, \dots, Z_d)$ on data rows with $X_1 \neq 0$, and use draws from the model to impute 0-entries. The completed data set is then used to train one of $m$ prediction models for $Y$. At prediction time, we draw one set of imputation (if the observed $X_1$ is 0) and $Y$ prediction from each of the model pairs and average the results to obtain the final prediction $\hat{Y}$. While imputing missing feature entries with draws from the posterior conditional distribution has the potential to alleviate bias introduced through under-reporting, it is not a priori clear how well the described method works when no missingness indicators are observed and true 0-entries are falsely imputed.

**Omission of rows** Omission of rows with missing feature entries provides a convenient complete case analysis setting. Similar to the multiple imputation scenario, excluding all missing values requires us to discard all data rows with 0-entries since true and false 0-entries are indistinguishable in the under-reporting setting. After discarding the respective rows, the remainder of the data can be used to train a prediction model for $Y$. If there is no model misspecification, e.g. in the linear case with $Y = \alpha + \beta^T Z$ where we train a linear model on observed features, training on only complete rows is guaranteed to asymptotically retrieve the true parameters $\hat{\alpha} = \alpha$ and $\hat{\beta} = \beta$ if $Z_1$ is not binary. Although differential under-reporting bias is successfully alleviated in estimation, under-reporting still affects outcomes at prediction time. If feature $Z_1$ is missing, the ground truth model $f(Z) = \alpha + \beta^T Z$ is not necessarily guaranteed to provide the most accurate (or fairest) prediction when applied to mismeasured features $X$.

### 3.7.2 Correction for under-reported features

Without missingness indicators, standard approaches for handling missing data based on omission and imputation are not guaranteed to diminish outcome disparities introduced through under-reporting without the introduction of additional biases. Instead, we propose missing data methods specifically tailored to the differential feature under-reporting setting. To achieve this, we separate the problem into two steps — estimation and prediction — which are both impacted by the under-reporting problem. For the estimation step, we provide a method that recovers the ground truth data generating model from observed data. For the prediction step, we derive the optimal group-dependent imputations for feature values observed as zero.

**Model estimation with augmented loss** We saw that training on rows with non-zero entries in the under-reported feature column recovers the truth model on $Z$ if the model is correctly specified and sufficient training data is available. In practice, some amount of misspecification is to be expected and discarding all observations with 0-entries can lower performance significantly if the training data set is small or contains a lot of correctly observed 0-entries. Instead, we propose an augmented or proxy loss. This proxy loss uses observed features $X$ to provide an unbiased estimate of the loss of a model $f$ on latent true features $Z$. The notion of employing unbiased estimators is widely acknowledged in stochastic optimization (Nemirovski et al., 2009) and has previously been used in the label noise setting (Natarajan et al., 2013).

Assume $Z \in \mathbb{R}^d$ has support $\mathcal{Z}$ and $y \in \mathbb{R}$ has support $\mathcal{Y}$. Let $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$ be a class of real-valued functions and $l : \mathcal{F} \times \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ be a bounded loss function. We assume missingness occurs only in the first feature, i.e. the observed feature vector $X = Z \odot \xi$ coincides with $Z$ in all but the first entry which is given by $X_1 = Z_1 \xi_1$ with $\xi_1 \sim \text{Bern}(m_1)$.

**Lemma 9** (Augmented loss)**.** *Assume fixed $f \in \mathcal{F}, z \in \mathcal{Z}$, $y \in \mathcal{Y}$ and $X \in \mathbb{R}^d$ defined by $X_1 = Z_1\xi_1$ and $X_{[2:d]} = z_{[2:d]}$. Define*

$$\tilde{l}(f, X, y) = \frac{1}{m_1} l(f, X, y) - \frac{1 - m_1}{m_1} l(f, [0, X_{[2:d]}]^T, y).$$

*If $Z \perp G$, we have that $\mathbb{E}_{\xi_1}\left[\tilde{l}(f, X, y)\right] = l(f, z, y)$.*

The fact that the augmented loss is unbiased with respect to under-reporting noise implies that a prediction model on observed data estimated with augmented loss, i.e.

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)} \left[ \tilde{l}(f, X, Y) \right],$$

asymptotically recovers the Bayes optimal model on the true features $Z$. If $Y = \alpha + \beta^T Z$, $\mathcal{F}$ is the class of linear functions $f : \mathbb{R}^d \to \mathbb{R}$, and $l(f, z, y) = (f(z) - y)^2$ denotes squared error loss, the true parameters $\hat{\alpha} = \alpha$ and $\hat{\beta} = \beta$ are retrieved. Note that squared error loss is not bounded and estimating $\hat{f}$ requires the additional constraint that $\tilde{l}(f, X, Y) \geq 0$.

Lemma 9 operates in a group-agnostic setting with $Z \perp G$. If feature distributions vary across groups and $X_1 = Z_1 \xi_1^g$ where $\xi_1^g \sim \text{Bern}(m_1^g)$ depends on group membership, the unbiased loss estimator takes the following form.

**Lemma 10** (Group-dependent augmented loss)**.** *Assume fixed $f \in \mathcal{F}, z \in \mathcal{Z}, y \in \mathcal{Y}, g \in \{0, 1\}$ and $X \in \mathbb{R}^d$ defined by $X_1 = Z_1 \xi_1^g$ and $X_{[2:d]} = z_{[2:d]}$. Define*

$$\tilde{l}(f, X, y, g) = \frac{1}{m_1^g} l(f, X, y) - \frac{1 - m_1^g}{m_1^g} l(f, [0, X_{[2:d]}]^T, y)$$

*Then, we have that $\mathbb{E}_{\xi_1^g} \left[ \tilde{l}(f, X, y, g) \right] = l(f, z, y)$.*

**Optimal prediction imputation value** Assume we are in the linear case with $Y = \alpha + \beta^T Z$ and we have access to the true parameters $\alpha$ and $\beta$, e.g. obtained via the augmented loss trick. Yet at prediction time, we only observe features with under-reporting $X$. What is the best possible prediction for an example of the form $x = [0, z_2, \ldots, z_d]$? Since $x_1 = 0$ could mean either that $z_1 = 0$ or that the entry is missing, it is intuitive that $\hat{y} = \alpha + \beta^T x$ does not minimize expected prediction error. Instead, we derive the optimal fixed prediction imputation value $x_1'$.

**Lemma 11** (Optimal prediction imputation value)**.** *Assume $f(Z) = \alpha + \beta^T Z$ is the ground truth model* *and $X$ a random vector of observed features. We set*

$$X' = \begin{cases} X \textit{ if } X_1 \neq 0, \\ [x'_1, X_{[2:d]}] \textit{if } X_1 = 0, \end{cases}$$

*where $x'_1$ is a fixed imputation value. Then, if $Z \perp G$,*

$$\arg\min_{x'_1} \mathbb{E}_X[(f(X') - Y)^2] = \mathbb{E}\left[Z_1 \mid X_1 = 0\right]$$

*is the optimal prediction imputation value.*

The Lemma shows that the loss-minimizing constant imputation value is the conditional mean of the true feature $Z_1$ given that the observed value is 0. This implies that, in alignment with earlier intuition, directly predicting with the observed $x = [0, z_2, \ldots, z_d]$ is sub-optimal except in cases in which in all observed 0-entries correctly observed. By law of iterated expectation and since $\mathbb{E}\left[Z_1 \mid X_1 \neq 0\right] = \mathbb{E}\left[X_1 \mid X_1 \neq 0\right]$, the optimal prediction imputation value in the setting of Lemma 11 can be expressed as

$$\mathbb{E}\left[Z_1 \mid X_1 = 0\right] = \frac{\frac{1}{m_1}\mathbb{E}\left[X_1\right] - P(X_1 \neq 0)\mathbb{E}\left[X_1 \mid X_1 \neq 0\right]}{P(X_1 = 0)},$$

which can be estimated directly from only observed data $X$ if the missingness rate $1 - m$ is known.

If feature distributions vary across groups and $X_1 = Z_1 \xi_1^g$ where $\xi^g \sim \text{Bern}(m_1^g)$ depends on group membership, Lemma 11 can be adapted as follows.

**Lemma 12** (Group-dependent optimal prediction imputation values)**.** *Assume $f(Z) = \alpha + \beta^T Z$ is the ground truth model, $X$ a random vector of observed features, and $G$ the group membership. We set*

$$X' = \begin{cases} X \textit{ if } X_1 \neq 0, \\ [x'^0_1, X_{[2:d]}] \textit{ if } X_1 = 0 \textit{ and } G = 0, \\ [x'^1_1, X_{[2:d]}] \textit{ if } X_1 = 0 \textit{ and } G = 1, \end{cases}$$

| Name | #Obs. | #Feat. | Groups | Binary outcomes |
|---|---|---|---|---|
| COMPAS (Angwin et al., 2016) | 7,214 | 6 | Race (51% African-American, 49% other), Gender (81% male, 19% female) | Two-year recidivism, violent recidivism |
| German credit (Repository, 1994) | 1,000 | 19 | Gender (69% male, 31% female) | Good credit |
| ACS Income (CA, 2018) (Ding et al., 2021) | 195,665 | 6 | Race (62% White, 38% other), Gender (53% male, 47% female) | Yearly income over $50,000 |
| Birth data | 39,365 | 51 | Medicaid (no 72%, yes 28%), Race (African-American 21%, other 79%) | Child placed in foster care within 3 years |

Table 3.2: Statistics of the data sets used in experiments. Data is split randomly into 80% for training and 20% for testing. For the first three data sets, we iterate over all outcome types, groups, and numerical features for missingness injection.

*where* $x_1'^0, x_1'^1$ *are group-dependent fixed imputation values. Then,*

$$\arg\min_{x_1'^g} \mathbb{E}_X \left[(f(X') - Y)^2\right] = \mathbb{E}\left[Z_1 \mid X_1 = 0, G = g\right]$$

*are the optimal group-dependent prediction imputation values for* $g \in \{0, 1\}$.

Similar to before, the optimal imputation values can be written as

$$\mathbb{E}\left[Z_1 \mid X_1 = 0, G = g\right] = \frac{\frac{1}{m_1^g}\mathbb{E}\left[X_1, G = g\right] - P(X_1 \neq 0 \mid G = g)\mathbb{E}\left[X_1 \mid X_1 \neq 0, G = g\right]}{P(X_1 = 0 \mid G = g)},$$

which can be estimated directly from observed data.

## 3.8 Experiments

In this section, we describe the data sets and experimental setup used for the empirical portion of our work. Experiments are conducted using several publicly available data sets from the algorithmic fairness literature, as well as one private county-level data set demonstrating the real-world relevance of differential feature under-reporting. We experiment with models fit directly on the mismeasured data as well as several solution approaches.

### 3.8.1 Data sets

**Publicly available data sets** We run experiments on three publicly available data sets as summarized in Table 3.2. Both the COMPAS data set (Angwin et al., 2016) and German credit data set (Repository, 1994) are widely used across the algorithmic fairness literature. The American Community Survey (ACS) Income data set is comprised of 2018 census data from California queried using the folktables package introduced by Ding et al. (2021). The data sets vary in size, number of features, and prediction tasks as shown in Table 3.2. We conduct all experiments with both gender and race as group columns if available. The group column under consideration for disparities is never included as predictive feature. Race information is never included as predictive feature.

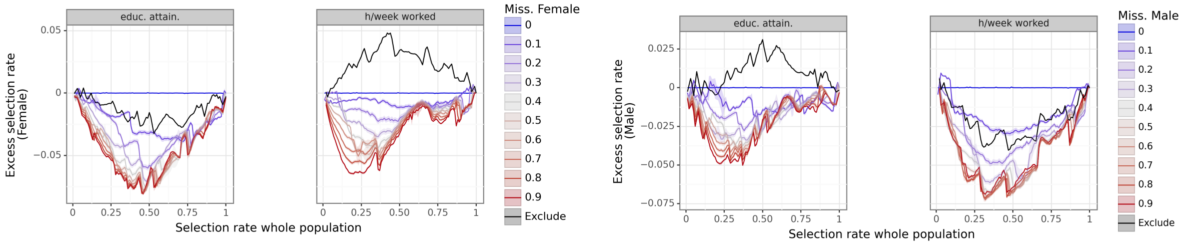
**Birth data** In addition to the publicly available data, we present here an analysis of a private administrative dataset we obtained from a County in the US with a rich administrative data system. The dataset contains information on newborn children and their families. The data contains demographics, child protective services history, birth record data, and mental and behavioral health information for those who used publicly funded services. We set up a prediction task that attempts to mirror the analysis described in the Hello Baby model methodology report from Allegheny County (for Social Data Analytics at the Auckland University of Technology, 2020). The Hello Baby model was developed to predict which families are at greatest risk of having their child removed by Child Protective Services (CPS) during their first three years of life, and is used to prioritize families for opt-in, voluntary supportive services. Using our data we train a similar model, and explore the effect of adding additional missingness to the behavioral and mental health data fields.

### 3.8.2 Setup

**Experiment stratification** Experiments are conducted on three publicly available data sets and one private administrative data set. For the publicly available data sets, we add artificial feature mismeasurement in the form of unobserved missingeness to one feature column in one group at a time and repeat the experiments for each available outcome column, group column, group within the column, and missingness level. Missingness levels range from 0-90% in 10 percentage point increments, and we add missingness to only one group at a time (e.g., we set 10% of a feature in the male group to 0 while leaving the features of the female group unchanged). Only numeric features are considered for unobserved missingness since, in administrative data, binary features are often categorical dummies or thresholded versions of underlying continuous count features. All models are trained with 80% of the data sets while withholding 20% for testing. Results are reported as averages over simulation runs on the test data.

**Semi-synthetic outcomes** Since all of the prediction tasks in the publicly available data sets are classification tasks with binary outcomes, we opt to generate semi-synthetic regression labels for our experiments. This is achieved as follows. We first fit a logistic regression model to the entire data set and extract the fitted probabilities. For the ACS Income data the probabilities are rescaled to more closely reflect the $50,000 income threshold. For the other data sets, we retain the probability predictions as they are. Next, we fit a linear regression model using the same features and the predicted probabilities as outputs. The fitted values from this linear model are chosen as the new "true" labels for our experiment. This outcome augmentation procedure allows us to generate artificial settings with a truly linear ground truth models to compare against similar to the setting studied in the theory portion of this work. At the same time, the feature covariance structures of the original data sets remain intact allowing us to obtain insights into what is likely to happen in real application settings. We further experiment with controlling the $R^2$ of the true linear model by adding additional random noise to the semi-synthetic outcomes and report implications for fairness outcomes in Appendix B.4.1.

**Solution approaches** We experiment with three common methods for handling missing feature values and our proposed method for correcting differential feature under-reporting bias. First, we explore how removing the entire feature column with mismeasurement impacts fairness outcomes. Second, we experiment with multiple imputation of all 0-entries. Crucially, this includes both missing values and values correctly observed as 0 since no distinction can be made between correctly and falsely observed zeros. As a third solution, we exclude all rows with 0-entries in the mismeasured feature from training, an approach that resembles the idea of a complete case analysis. Similar to the imputation approach, removing all rows with 0-entries rather than only the rows in which the value is actually missing is necessary since no missingness indicators are observed. Lastly, we employ our proposed method and fit models using the group-dependent augmented loss introduced in Section 3.7.2. Predictions are made with group-dependent optimal prediction imputation values for observed 0-entries. All solution approaches are compared against



(a) Female group

(b) Male group

Figure 3.2: Excess selection rate of group at different selection rates of the whole population with synthetic outcomes using the ACS Income data set. Each panel represents a feature that has been corrupted by missingess in independent runs of the experiment. Feature missingness is added to the same group with 0-90% missing in 10 percentage point increments. The black curves show performance when excluding the whole feature column from modeling. Results are reported as averages over 50 runs on the test set. Shaded areas correspond to one standard deviation in each direction of the mean.

the fairness outcomes of the models trained on differentially unobserved data.

**Experiment setup for birth data** We conduct our experiments on the birth data separately from the procedure used on the publicly available data sets in order to showcase a real example of the problem of unobserved feature missingness. As before, the data is separated into 80% for training and 20% for testing. Without any augmentation of the target labels in the data set, we fit three separate logistic regression models. (1) A logistic regression model using all of the available features. (2) A logistic regression model using all of the features after behavioral health related columns were set to 0 for all individuals, i.e. mothers, that are not insured through Medicaid. (3) A logistic regression model on only the feature columns that are not related to behavioral health information. Results of the the three prediction models on the test data set are stratified both by whether individuals are covered by Medicaid or not and by whether individuals identify as African-American or not for the purpose of illustration. Neither Medicaid nor racial information are used as features in any of the models.

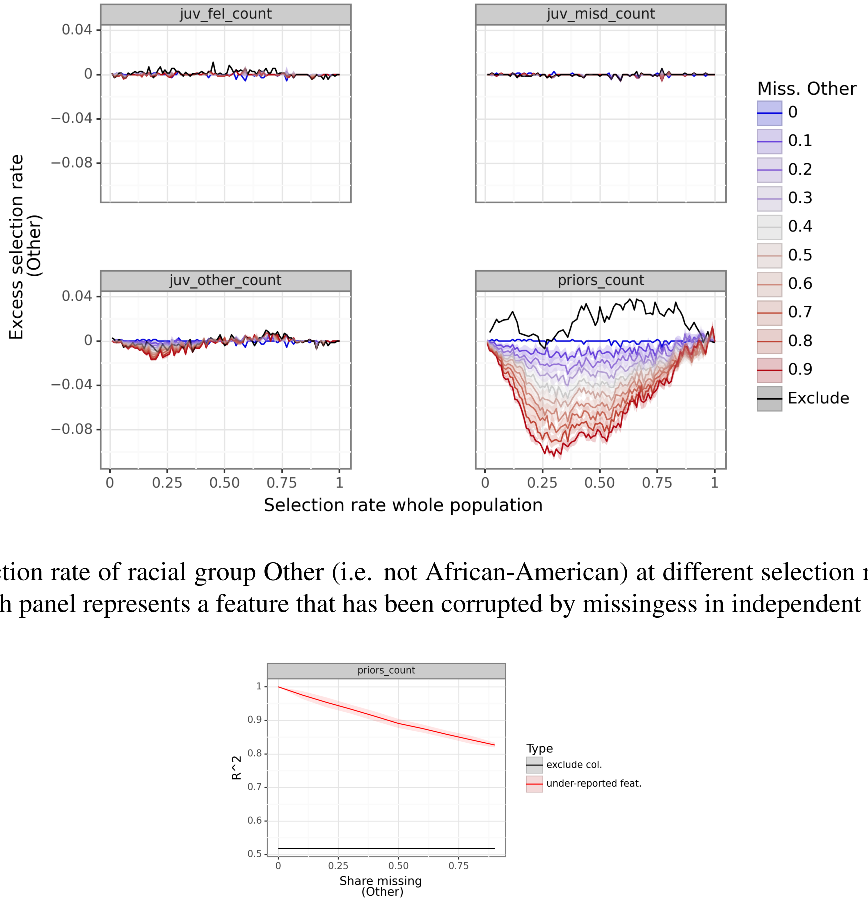
## 3.9 Empirical results

### 3.9.1 The impact of unobserved feature missingness

**ACS Income data** A subset of results for the experiments on the ACS Income data are displayed in Figure 3.2. We see that addition of unobserved feature missingness to the features ‘education attainment’ and ‘hours worked per week’ consistently leads to under-selection of the group with missingness. This is true irrespective of whether feature missingness is injected into the female sub-group of the population or the male sub-group, and we observe the same effect when missingness is added based on the individuals’ racial group instead of gender. The figure additionally suggests that more missingness generally leads to increasing under-selection for all displayed variables. Intuitively, it makes sense that both education attainment and hours worked per week contribute positively to predicted income which is confirmed by the parameter estimates of our models. Exploration of the covariance matrix of the unbiased features further reveals that all numeric columns in the data set are positively correlated which together creates a setting reminiscent of the Case 2 scenario studied in Section 3.6. On a high-level, the scenario predicts that the group with more missingness is under-selected in the given setting which aligns with our observations. In addition to selection rate disparity, feature under-reporting in the data also leads to decreased model accuracy as displayed in Figure B.1. The parameter estimates in Figure B.2 display an attenuation effect as predicted in Section 3.5.

**COMPAS data** We display excess selection rates for missingness in the racial group Other (i.e. not African-American) for a subset of features in Figure 3.3. While the results presented here concentrate on unobserved missingness in count features, we discuss the case of missingness in the feature ‘age’ in Appendix B.4.2. The feature ‘priors count’, i.e. the number of previous criminal offenses individuals have been convicted of, emerges as an important feature with respect to missingness. We clearly see that missingness in priors count leads to under-selection of the group with missingness. This pattern repeats itself for any of the groups with missingness, and both of the available prediction outcomes. The more data is missing from a group, the larger the occurring outcome disparity. Similarly to the previous results, this suggests a setting of Case 2 as discussed in Section 3.6. As before, the parameter estimates under unobserved feature missingness suggest an attenuation effect which is displayed for the feature ‘priors count’ in the left column of Figure 3.4. As missingness is injected into ‘priors count’, the model shifts weight from ‘priors count’ to the positively correlated count features by increasing the respective parameter estimates.

Missingness in the number of previous criminal offenses could be interpreted as an extreme case of crimes that do not result in arrest. Assuming that one demographic group is more likely to be convicted for



(a) Excess selection rate of racial group Other (i.e. not African-American) at different selection rates of the whole population. Each panel represents a feature that has been corrupted by missingess in independent runs of the experiment.

(b) Test set $R^2$ of models with missingness and model with excluded column using priors count as example.

Figure 3.3: Results for experiments on COMPAS data set with synthetic two-year recidivism outcomes. Feature missingness is added to the Other group with 0-90% missing in 10 percentage point increments. The black curves show performance when excluding the whole feature column from modeling. Results are reported as averages over 30 runs on a test set. Shaded areas correspond to one standard deviation in each direction of the mean.

committed crimes than the other group, the result implies that the same already more frequently targeted group may additionally be flagged as high risk for recidivism at disproportionately high rates. Racial disparities in arrest rates and police encounters more generally are well-documented in the US (Alexander, 2020; Butcher et al., 2022; Fogliato et al., 2021; Pierson et al., 2020) which highlights the importance of this observation.

**German credit data** Our experiments suggest that addition of unobserved feature missingness to one of the two gender groups in the German credit data set has only marginal fairness implications. Figure B.4 depicts the results for synthetic outcomes and addition of different amounts of missingness to the features of the male group. We can see that, for any of the considered features, the amount of missingness injected has little to no effect on the excess selection rate of the male group. However, when selecting rates of around 10-15% from the whole population *any* amount of missingness in the installment feature appears to results in a slight over-selection of the male group. The installment feature in the German credit data set is discretized into four values with lower values indicating a higher installment rate. Incorrectly observed 0-values may thus suggest a high installment rate which is indicative of good credit.

### 3.9.2 Standard approaches for handling missing data

**Removing entire feature columns** We return to the ACS Income data results depicted in Figure 3.2 in order to examine the unfairness mitigation potential of dropping the entire feature column when features are impacted by unobserved missingness. In the left panel of the figure (female group), a model without the feature 'education attainment' outperforms the models on differentially observed data in terms of fairness at every missingness level. However, for the feature 'hours worked per week', the disparity is not diminished and instead flipped in sign. While training on data with missingness leads to a setting in which missing 'hours worked per week' entries lead to under-selection, removing the feature column entirely leads to over-selection of the female group. This over-selection occurs because female individuals in the data set report to work on average less than male individuals (35.43h/week vs. 40.05h/week), and the true and estimated regression parameters of the feature are positive for all missingness levels (compare Figure B.2). Removing the feature entirely hides this difference hence benefiting the female group. The reverse argument applies to the feature 'education attainment' which is slightly increased in the female group, and the picture of disparity when removing entire features is flipped when comparing to the right side of Figure 3.2 (male group).

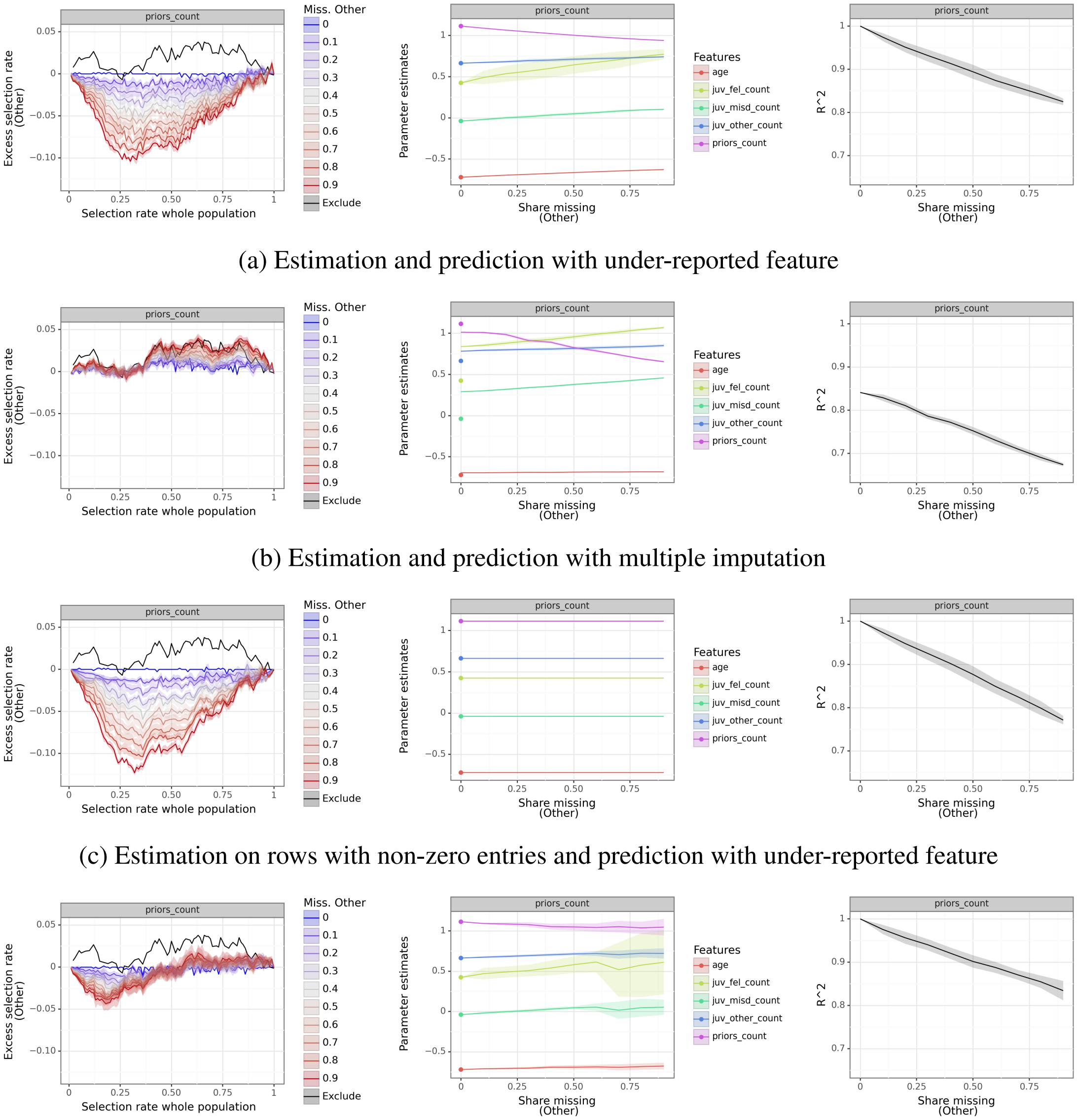
A similar effect occurs when removing the feature ‘priors count’ in the experiments on the COMPAS data set. Figures 3.3 and 3.4 show that missingness in the feature for the group Other (i.e. not African-American) leads to under-selection of the Other group while removal of the entire feature column leads to over-selection of the group. This phenomenon occurs because defendants in group Other in the data on average have fewer priors (2.46) than the African-American defendants (4.44), and the true and estimated parameters of ‘priors count’ are positive at every missingness level.

Taken together, these findings illustrate that removing feature columns with unobserved missingness entirely does not necessarily mitigate selection rate disparities. In fact, the selection rate disparities can be reversed by excludig the columns altogether.

**Multiple imputation** Figures 3.4 and B.3 display the results of multiple imputation experiments on the COMPAS data set assuming the same features and groups as previously. We see that the excess selection rate has flipped signs, and instead of being under-selected, the group with more missingness in the feature ‘priors count’ is over-selected after multiple imputation. On a high-level, this is because the ‘priors count’ feature naturally has a lot of true 0-entries which are wrongfully imputed as positive values in this setting. These wrong imputations lead to considerably decreased test set $R^2$ of the model, and introduce bias in the parameter estimation as well as the prediction step. In comparison to training on mismeasured features directly, the bias in parameter estimates with multiple imputation is considerable even for small amounts of missingess or no missingness at all (Figure 3.4).

Figure 3.4 further suggests that, for large amounts of missingness, the excess selection rate under multiple imputation follows a similar pattern as the excess selection rate when removing the feature ‘priors count’ altogether. This is unsurprising since imputation is conducted using the features already present in the model. If close to no true values of ‘priors count’ are observed, an imputed version of the feature column contributes little to no additional information to the model trained on the same features used for imputation.

Overall, the results demonstrate that multiple imputation in the unobserved feature missingness setting does not necessarily lead to more equitable outcomes. In fact, the cost incurred by wrongfully imputing



(a) Estimation and prediction with under-reported feature

(b) Estimation and prediction with multiple imputation

(c) Estimation on rows with non-zero entries and prediction with under-reported feature

(d) Estimation with group-dependent augmented loss and prediction with optimal group-dependent imputation (**our method**)

Figure 3.4: Excess selection rates of group Other (i.e. not African-American) (left columns), parameter estimates (middle column), and test set $R^2$ (right columns) when missingness is injected into the feature 'priors count' in group Other using the COMPAS data set and synthetic two-year recidivism outcomes. In (a), the model is trained and evaluated using the mismeasured feature directly. For (b), we first train a multiple imputation model and then train and evaluate the prediction model using probabilistic imputations. For (c), the model is trained on only rows without 0-entries in 'priors count' and evaluated on the mismeasured data. In (d), we train with group-dependent augmented loss and use group-dependent optimal imputation values for prediction. Results are reported as averages over 30 runs. Shaded areas correspond to one standard deviation. The solid dots in the middle column correspond to the true parameters from a semi-synthetic ground-truth model. Note that in order to preserve readability, parameter estimates are only displayed for continuous features. The models additionally use sex and the categorical charge degree as features. Figure B.6 provides an overlay plot of the rightmost column for easy comparison.

true 0-entries can exceed the benefit of imputing the missing values. In addition, there are severe ethical concerns around using imputation strategies in individual-level high-stakes applications like recidivism prediction. Imputation has the potential of inflicting real harm as discussed in previous literature (Ahmad et al., 2019).

**Removing rows with 0-entries** Figures 3.4 and B.5 summarize the results of training only on rows without 0-entries in the mismeasured feature on the COMPAS data set. We note that unobserved feature missingness does not lead to estimation error in the parameter estimate of 'priors count' in this setting (Figure 3.4). This is because (1) removing the rows with 0-entries removes all mismeasurement from the feature, and (2) the linear model is well-specified in our semi-synthetic experimental setup.

Although the solution approach successfully recovers the correct model for the latent correctly measured features, the feature missingness still introduces bias into the system via the prediction step. In fact, we observe that the selection rate disparity is increased, i.e. the excess selection rate of the group Other is larger in magnitude, as compared to the model trained on mismeasured data directly. While the model trained on mismeasured data is able to shift weight from 'priors count' to other correlated count features as more and more entries for 'priors count' are missing, the model trained on only rows with non-zero entries cannot make use of the feature correlations at prediction time which ultimately leads to increasing rather than decreasing disparities. With the same reasoning, the test set performance as measured by $R^2$ is decreased when omitting all rows with 0-entries as displayed in the figure.

### 3.9.3 Correction for under-reported features

We contrast the performance of our method and standard approaches for handling missing features at the example of the 'priors count' feature in the COMPAS data set. The results in Figure 3.4 show that selection rate disparities decrease considerably when estimating the model with group-dependent augmented loss and using group-dependent optimal imputation for observed 0-entries at prediction time. In contrast to the multiple imputation results, this fairness improvement comes at no visible cost in performance. In fact, the average test set $R^2$ of the corrected model is very similar to, and even slightly higher than, the test set $R^2$ of the model trained directly on under-reported data (see Figure B.6). Despite some variability, the average parameter estimates of the corrected model appear more stable across different amounts of

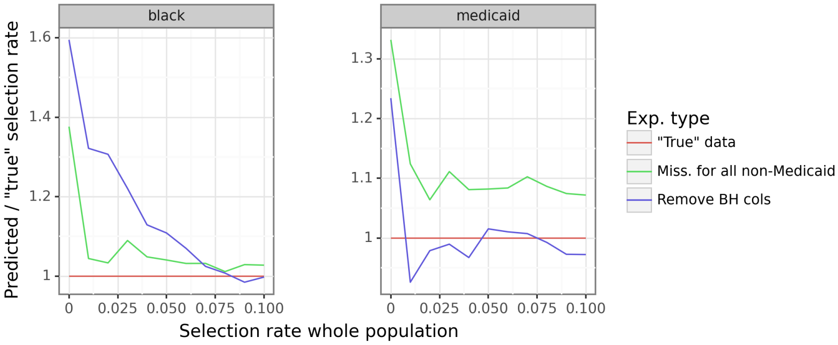


Figure 3.5: Fraction of predicted selection rates of different models and the "true" data selection rate for the birth data set. On the left, the results are displayed for the sub-population of Black individuals, on the right, the results are displayed for the sub-population that is insured through Medicaid. The selection rate of the whole population is considered to be 10% or lower which reflects a realistic range for this predictive risk modelling setting.

missingness than in the mismeasured feature and multiple imputation models which suggests that the method successfully diminishes the bias under-reporting introduces into model estimation.

### 3.9.4 Results on the birth data

Our experiments on the birth data set suggest that under-reporting of all behavioral health data for the non-Medicaid population leads to over-selection of the Medicaid population (see blue line in Figure 3.5). In particular, if the selection rate of the whole population is 1.3-10%, the Medicaid population is selected about 10% more often than in the "true" data setting. Note that in reality this difference could be even larger because the "true" data features were likely already differentially missing. As shown in the Figure, some of the resulting disadvantage is still observable when evaluating performance in the group of Black individuals. This can be explained by the fact that the two group variables are positively correlated in the data set ($\rho = 0.44$).

Lastly, we observe that removing the behavioral health columns altogether leads to a reversal of selection rate disparities in the Medicaid / non-Medicaid groups, and significantly increases selection of the Black sub-population at population selection rates of less that about 7% (pink line). We conclude that removal of the under-reported columns is not a successful remedy for the disparities caused by differential under-reporting.

## 3.10 Discussion

### 3.10.1 Main findings

**Impact of differential feature under-reporting** We study the impact of differential feature under-reporting on algorithmic fairness by (1) presenting and examining an analytically tractable model of unobserved feature missingness, and (2) empirically exploring performance under hidden missingness on semi-synthetic and real-world data. Our work assumes a thresholded regression setting inspired by predictive risk modeling in the public sector. We demonstrate how feature missingness in the studied setting impacts prediction outcomes of the downstream model in two ways. First, the missingness leads to misestimation of model parameters as compared to a model with access to correctly observed data (estimation step). Second, observations with missing features have different predictions than their counterfactual counterparts with fully observed features (prediction step).

**Estimation bias** Assuming outcomes are a linear function of the latent correctly measured features, our results suggest an attenuation effect of missingness on the respective parameter in linear regression. This is in line with previous observations from the additive feature noise case (Fuller, 1987; Hausman, 2001). Instead of fully weighing the mismeasured feature, the model with missingness shifts weight to fully-observed features that are correlated with the mismeasured feature. Here, 'shifting weight' can imply an increase or decrease in another feature's estimated model parameter depending on the signs of feature correlations and ground-truth parameters. At a high level, our results suggest that the parameter estimate for a fully observed feature increases if the correlation to the feature with missingness and the ground-truth paramater for the feature with missingness point into the same direction. Otherwise the parameter estimate is decreasing. While these theoretical results require some simplifying assumptions, our experiments confirm both the occurrence of attenuation bias as well as a shift of weight to other correlated features under real world feature and group dependence structures.

**Fairness in selection rates** We study the equity of predictions under missingness by tracking how the selection rate disparity between groups changes when group-dependent levels of unobserved feature missingess are introduced. Since we are studying a setting with two groups and the selection rate of the whole population is kept fixed, e.g. the top 2% are selected irrespective of whether features are mismeasured or fully observed, an increase in the selection rate for one of the groups directly implies a decrease in the selection rate for the other group. In turn, it is sufficient to argue about the excess selection rate for one of the groups to make conclusions about selection rate disparities. Under a set of simplifying assumptions, we analytically show that, when selecting small shares of the population, the group with more unobserved feature missingness is under-selected as compared to the ground truth model if and only the variance of predictions is decreased by missingness. In this case, the selection rate disparity between groups increases with increased missingness if the group with more missing values coincides with less selected group in the ground-truth model. However, our work also shows that, analytically, the reverse can occur and disparities are decreased when the features of the less selected group are impacted by missingness. We study the conditions for both cases in a jointly Gaussian feature setting and find that the direction of the effect depends on first and second moments of the features, their correlations, and the ground-truth parameters. Although both scenarios can occur in theory, the observations from our experiments suggest that, typically, unobserved feature missingness leads to increasing selection rate disparities. The only empirical example we find for decreasing disparities is discussed in Appendix B.4.2.

**Failure of standard missing data methods** Omission of columns or rows with missingness and imputation of missing values present some of the most common strategies to overcome problems of missing data. We experiment with these mitigation strategies while exploring their impact on fairness outcomes. Our results demonstrate that omission of features can lead to either increases or decreases in selection rate disparity depending on the exact setting. This finding is in agreement with previous lines of work that study how removing seemingly unfair features can lead to outcomes that are more inequitable than the outcomes of models including the features (Garg et al., 2020; Bartik and Nelson, 2019; Khani and Liang, 2021).

Multiple imputation in our setting has the additional difficulty that no missingness indicators are observed. Instead of imputing exclusively missing values, we can only experiment with imputing all feature values marked as 0 which we observe to lead to an increase in outcome disparities in our setting. Intuitively, the cost incurred by wrongfully imputing correctly observed 0-entries can outweigh the benefits of imputing missing values, particularly when true 0-entries are common.

Omission of rows with missingness suffers from the same problem as imputation—no indicators for missingness are observed. We instead experiment with removing all rows in which the feature *could* be missing. For well-specified models, this allows us to recover the true model on the correctly measured features. However, this model is not necessarily optimal for the mismeasured features available at prediction time. In fact, we observe that omission of rows with 0-entries leads to increased disparities rather than decreased disparities as compared to training on the mismeasured data directly.

**New correction method for under-reported features and future directions** We propose a new method for handling missing data that is specifically tailored to the setting of differentially under-reported features. The method is separated into two steps — model estimation and prediction — which are both impacted by the bias introduced through feature mismeasurement. For the estimation step, we propose a loss augmentation that provides an unbiased estimate of the model loss with respect to the latent true features. For prediction, we derive the set of group-dependent fixed imputation values that minimize expected prediction error. Our experimental results suggest that the method successfully reduces selection rate disparity introduced by feature under-reporting with little to no loss in performance as measured by test data $R^2$.

The proposed method requires knowledge of the rate of missingness in the under-reported feature column. In some applications, it may be possible to obtain estimates of the missingness rate from separate validation data. In future work, we plan to focus on estimation of under-reporting rates from the under-reported data directly.

### 3.10.2 Practical implications

Differential feature under-reporting in the form of differential unobserved missingness is a common phenomenon in administrative data settings. Data records are generally more complete for individuals who rely more consistently on public services (e.g. in the form of public health coverage). In many predictive risk assessment settings, the segment of the population with more complete observations overlaps with sub-populations that are more commonly selected as high-risk. Critics have argued that differential data availability is a key driver of the observed disparity in selection rates. When being classified as high risk (be it ”correctly” or ”incorrectly”) subjects one to greater scrunity of burden, this may disadvantage those with more complete data (Eubanks, 2018). Overall, the results of our study lends further credence to the concern by demonstrating how unobserved feature missingness generally leads to under-selection of a group that is already less frequently identified as high-risk. While, as we demonstrate, it is theoretically possible for groups with greater data availability to be under-selected, the feature dependence structure under which this occurs appears uncommon in practice.

We illustrate the increased selection rate that individuals who rely on public healthcare coverage may experience in the context of a real-world risk assessment problem. Following the idea of Allegheny County's Hello Baby program (for Social Data Analytics at the Auckland University of Technology, 2020), we build a model that predicts the risk that a newborn child will be removed from their family by Child Protective Services (CPS) within three years based on county-level data. The data set contains behavioral and mental health information on the parents which can be assumed to be more complete for families that rely on public insurance. We note that, for privately insured individuals, some of this information may still be observed, e.g. because the individual was publicly insured previously, or individual information has been collected explicitly, but a lot of the information can be assumed missing. Our experiments suggest that further missingness in behavioral and mental health related information for the privately insured sub-population leads to an increase in high-risk predictions for the publicly insured group. We hypothesize that this effect would be even larger if the data set was not already missing large portions of the feature observations for the privately insured group. This finding implies an unfair targeting of the publicly insured sub-population as high-risk. Since African-American mothers in the studied data set are publicly insured more frequently than mothers from other racial groups, these results also suggest that the newborns of African-American mothers are predicted to be at high-risk of being removed from their family by CPS at unfairly inflated rates. Of course, depending on the type of intervention, the inflated risk may lead to an advantage or a disadvantage for the families. In the Hello Baby setting, high risk is tied to eligibility for voluntary supportive services provided by external county-funded service providers.

Our work proposes a technical remedy for the impact of differential feature under-reporting as a driver of disparities in selection rates. While standard missing data methods did not lead to more equitable outcomes in general, our experiments with semi-synthetic data suggest that this new method can reduce disparities considerably with little to no decrease in model accuracy. The applicability and performance of this approach in administrative data settings like the Hello baby program remains an interesting and important avenue for future work.

# Chapter 4

# Sandbox tool to (bias)stress-test fairness algorithms

Based on (Akpinar et al., 2022b): Nil-Jana Akpinar, Manish Nagireddy, Logan Stapleton, Hao-Fei Cheng, Haiyi Zhu, Steven Wu, and Hoda Heidari. A sandbox tool to bias(stress)-test fairness algorithms, *working paper / preprint arXiv:2204.10233*.

## 4.1 Introduction

Machine Learning (ML) increasingly makes or informs high-stakes decisions allocating or withholding vital resources to individuals and communities in domains such as employment, credit lending, education, welfare benefits, and beyond. If not done carefully, ML-based decision-making systems may worsen existing inequities and impose disparate harms on already underserved individuals and social groups. This realization has motivated an active area of research into quantifying and guaranteeing fairness for ML. Prior work has proposed various mathematical formulations of (un)fairness as predictive (dis)parities (Berk et al., 2018; Dwork et al., 2012; Hardt et al., 2016; Joseph et al., 2016) and fairness-enhancing algorithms to guarantee the respective parity conditions in the trained model's predictions (Agarwal et al., 2018a; Hardt et al., 2016; Calmon et al., 2017; Feldman et al., 2015; Zhang et al., 2018; Kamiran et al., 2012). Our work argues that these interventions are not sufficiently well-understood to warrant practical uptake. One crucial limitations of these algorithms is the fact that they are agnostic to the underlying *sources* of the observed unfairness. As a result, applying them in practice may simply hide the real problem by ensuring narrowly defined notions of parity in predictions. As a result, what these methods seemingly gain in observational parity can come at the cost of predictive disparity and accuracy loss in deployment, and at worst, they can become an instrument of fair-washing (Aïvodji et al., 2019).

As an example, consider a hypothetical healthcare setting in which electronic healthcare data is used to determine which patients are selected for a specialized treatment. Assume the hospital wants to ensure their prediction model is fair across a demographic majority and minority group. In this setting, it may seem intuitive to promote fairness by enforcing an Equalized Odds constraint. Equalized Odds ensures that, given the true need for the procedure is the same, the model's decision to select a patient is independent of the patient's group membership. While this may seem like a viable solution to combat unfairness, it is agnostic to the types of data bias that cause outcome disparities. Training data in healthcare prediction tasks like this is often plagued by biases (Obermeyer et al., 2019; Chen et al., 2021). In our example, we may not have access to patients' true need for the procedure and instead default to healthcare cost as a proxy outcome to train the model. Since access to healthcare has historically been lower for some minority groups, this can lead to a setting in which minority group patients selected for the procedure are sicker than their majority group counterparts even when enforcing Equalized Odds. Blindly applying an off-the-shelf Equalized Odds fairness enhancing method without understanding the types of bias present in this setting could thus hide the real problem while creating an illusion of fairness.

Our work aims to address the above shortcoming by offering a simulation framework for examining fairness interventions in the presence of various biases. This offers an initial yet crucial step toward a broader research agenda: to trace the limitations and scope of applicability of fairness-enhancing algorithms. We start with the observation that the ML pipeline consists of numerous steps, and distinct types of biases (e.g., under-/over-representation of certain groups, label bias, or measurement bias in the training data) can creep into it at various stages, amplifying or concealing each other in the trained model's predictive disparities. The fair-ML scholarship currently lacks a comprehensive framework for specifying the conditions under which each algorithmic fairness-enhancing mechanism effectively removes specific types of biases—instead of simply covering up their manifestations as unfairness. For example, it is unclear what type of intervention (e.g., pre-, in-, or post-processing) one must employ depending on the underlying cause of the observed statistical disparity. As a concrete instance, a careful investigation of the relationship between biases and fairness remedies may reveal that if the source of unfairness is label bias among examples belonging to the disadvantaged group, imposing fairness constraints on ERM may be more effective than certain types of pre-processing or post-processing techniques. The reverse might be true for a different bias (e.g., biased choice of hypothesis class).

**Our simulation tool**. Motivated by the above account, in this work, we identify and simulate various (stylized) forms of bias that can infiltrate the ML pipeline and lead to observational unfairness. We prototype a sandbox toolkit designed to facilitate simulating and assessing the effectiveness of algorithmic fairness methods in alleviating specific types of bias, by providing a controlled environment. We call this process the *bias(stress)-testing* of algorithmic interventions. Our sandbox offers users a simulation environment to stress-test existing remedies by

1. simulating/injecting various types of biases (e.g., representation bias, measurement bias, omitted variable bias, model validity discrepancies) into their ML pipeline;

2. observing the interactions of these biases with one another via the predictions produced at the end of the ML pipeline (i.e., through the trained model);

3. and testing the effectiveness of a given algorithmic fairness intervention in alleviating the injected biases.

This chapter offers a preliminary implementation of the idea (see footnote 1) along with a detailed proof-of-concept analysis showing its utility. The sandbox is currently realized as a python library and we are working to add a visual user interface component in the future. We emphasize that the tool needs to be further developed and thoroughly evaluated before it is ready to be utilized beyond educational and research settings. The current implementation can be utilized

- in research settings to explore the relationships between bias and unfairness, and shape informed hypotheses for further theoretical and empirical investigations;

- as an educational tool to demonstrate the nuanced sources of unfairness, and grasp the limitations of fairness-enhancing algorithms;

Ultimately once the tool is fully developed and validated, we hope that it can be utilized by practitioners interested in exploring the potential effect of various algorithmic interventions in their real-world use cases. This will be an appropriate usage of the tool *if* (and this is an crucial if) the bias patterns in the real-world data are well-understood.

**Counterfactual comparisons**. The key idea that distinguishes our tool from existing ones is the possibility of evaluating fairness interventions beyond observational measures of predictive disparity. In particular, we can test whether a given remedy can alleviate the injected bias by comparing the predictions resulting from the intervention in the biased setting with the true labels *before* bias injection. This ability to compare with the unbiased data provides an ideal baseline for assessing the efficacy of a given remedy. We note, however, that the viability of this approach requires access to unbiased data. We, therefore, strongly recommend restricting the use of our tool to *synthetic* data sets—unless the user has a comprehensive and in-depth understanding of various biases in the real-world dataset they plan to experiment with.

**Remark.** *Note that in the case of real-world applications, one can rarely assume the training data are free of bias. However, if the practitioner is aware of what biases are present in the data (e.g., under-representation of a specific group) our toolkit may still allow them to obtain practically relevant insights concerning the effect of their fairness interventions of choice (e.g., up-sampling) on alleviating that bias—* assuming *that we can extrapolate the observed relationship between the amount of* additional *bias injected and the trained model's unfairness. We leave a thorough assessment of our toolkit's applicability to real-world data as a critical direction for future work.*

**Case study**. We demonstrate the utility of our proposed simulation tool through a case study. In particular, we simulate the setting studied in (Blum and Stangl, 2020). Blum and Stangl offer one of the few rigorous analyses of fairness algorithms under specific bias conditions. Their work establishes an intriguing theoretical result that calls into question conventional wisdom about the existence of tradeoffs between accuracy and fairness. In particular, their theoretical analysis shows that when underrepresentation bias is present in the training data, constraining Empirical Risk Minimization (ERM) with Equalized Odds (EO) conditions can recover a Bayes optimal classifier under certain conditions. Utilizing our tool, we investigate the extent to which these findings remain valid in the finite data case. Our findings suggest that, even for relatively simple regression models, a considerable amount of training data is required to recover from under-representation bias. In the studied settings with smaller data sets and little to moderate under-representation bias the intervention model showed to be no more successful in recovering the Bayes optimal model than a model without intervention. We then investigate the effectiveness of the in-processing EO intervention when alternative types of biases are injected into the training data. We observe

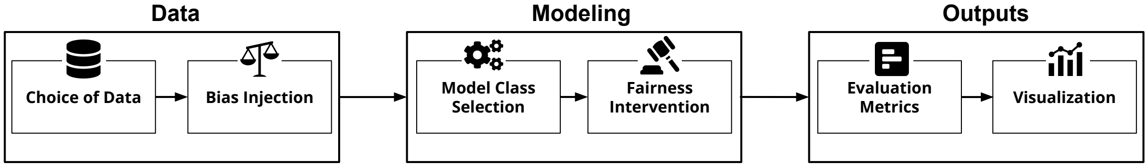


Figure 4.1: Flowchart illustrating the modules of the sandbox framework.

that the intervention model struggles to recover from the other studied types of biases. In some of the bias settings, such as a difference in base rates or differential label noise, the model with Equalized Odds intervention can provably not recover the Bayes optimal classifier since the latter does not fulfill Equalized Odds. Finally, we contrast the in-processing approach with the original post-processing method introduced by (Hardt et al., 2016) to ensure EO. As we discuss in Sections 4.3 and 4.4, our empirical analysis identifies several critical limitations of these methods.

**Remark.** *Our proof-of-concept demonstration deliberately addresses a small number of fairness-enhancing algorithms, and evaluates them in the presence of a wide range of data biases. While the sandbox can be used to contrast a wide array of fairness definitions and algorithms, such an analysis is beyond the scope of the current contribution and is left as an important avenue for future work.*

In summary, we present a first implementation of the unified toolkit to bias(stress)-test existing fairness algorithms by assessing their performance under carefully injected biases. As demonstrated by our case study, our sandbox environment can offer practically relevant insights to users and present researchers with hypotheses for further investigations. Moreover, our work provides a potential hands-on educational tool to learn about the relationship between data bias and unfairness as a part of the AI ethics curricula. Once evaluated and validated carefully, we hope that this tool contributes to educating current and future AI experts about the potential of technological work for producing or amplifying social disparities, and in the process, impacting lives and society.

## 4.2 Description of the Sandbox

The proposed sandbox presents a tool to understand the effectiveness of fairness enhancing algorithms under counterfactually injected bias in binary classification settings. The tool can be visualized as a simplified ML pipeline, with room for customization at each step. As indicated by its name, the sandbox prioritizes modularity and allows users to play around and experiment with alternatives at every stage. We summarize the six stages of the pipeline, which are illustrated in Figure 4.1, in the following. Note that, at the current time, some of implementation details as well as a visual user interface are still under development.

1. **Choice of Data:** The sandbox will allow users to select one of three options: input their own dataset, select one of the benchmark datasets in fair ML (e.g. Adult Income (Kohavi and Becker, 2017)), or synthetically generate a dataset. For custom data, the user will be asked to indicate which columns are to be used as group, outcome and feature columns. For synthetic data, which is the recommended option at this time, we provide a rigorous helper file which allows users to customize how the dataset is built. For example, we permit users to determine the number and quality (e.g. categorical or numeric) of features, the distribution of values for each feature, and the proportion of examples in different protected groups. The protected attribute is assumed to be binary. In addition, users can choose how labels are generated where label distributions are allowed to vary across groups. If desired, data can be sampled from a causal graph as demonstrated in Appendix C.3.

2. **Bias Injection:** The crux of our sandbox pipeline is the injection of different types of biases. In this iteration of the tool, we provide the options to inject representation bias, measurement bias, sampling bias, and label bias (Mehrabi et al., 2021; Frénay and Verleysen, 2013) which spans a large portion of the bias types discussed in the fair ML literature. Support for other types of bias will be added in the near future. In addition to injecting bias into the whole data set, the sandbox tool allows for application of biases at the intersection of protected attributes and other variables. For example, users can decide to under-sample only the positively-labeled examples from a group. Users are able to inject multiple biases at once which allows for realistic bias patterns that are multi-faceted in nature.

3. **Model Class Selection:** The proposed sandbox tool is compatible with any machine learning model

in the scikit-learn paradigm. We encourage the use of the so-called white-box classifiers, as they allow for greater ease when reasoning about the results obtained throughout the sandbox pipeline and present use cases of the sandbox with logistic regression in Sections 4.3 and 4.4.

4. **Fairness Intervention:** We make use of four fairness enhancing algorithms from the Fairlearn package (Bird et al., 2020) covering pre-processing, in-processing and post-processing techniques. First, the `CorrelationRemover` pre-processing algorithm filters out correlation between the sensitive feature and other features in the data. Next, the `ExponentiatedGradient` and `GridSearch` in-processing algorithms operate on a model and are based on Agarwal et al. (2018a). Finally, the `ThresholdOptimizer` post-processing algorithm adjusts a classifier's predictions to satisfy a specific fairness constraint.

   Possible fairness metrics for in- and post-processing algorithms are Equalized Odds, Equality of Opportunity, Demographic Parity, Error Rate Parity, and False Positive Rate Parity. For example, in Section 4.3, we utilize the `GridSearch` algorithm subject to an Equalized Odds constraint. In Appendix C.3 we consider Equalized Odds, Equality of Opportunity and Demographic Parity metrics.

5. **Evaluation Metrics:** Any scikit-learn supported machine learning performance metric for classification can be utilized in our sandbox framework. Examples include precision, accuracy, recall, F1 score, etc. Additionally, the sandbox also supports fairness metrics for evaluation, such as Equalized Odds or Demographic Parity disparities. For example, we obtain Equalized Odds disparities for the demonstrations provided in Sections 4.3 and 4.4.

6. **Visualization:** The sandbox tool outputs several figures including a visualization of the effectiveness of a fairness intervention at dealing with a particular type of bias. We note that various notions of performance are supported including more traditional measures of performance such as accuracy. Figure 4.2 provides an example visualization output of the sandbox. The figure displays the performance of a learned model in the selected metric (here, accuracy) over different degrees of bias (here, under-sampling examples from one group). Our sandbox allows us to compare performance in two dimensions: (1) Between models with and without a fairness intervention, and (2) On biased data versus unbiased ground truth data. In the figure, we show the latter comparison. We inject

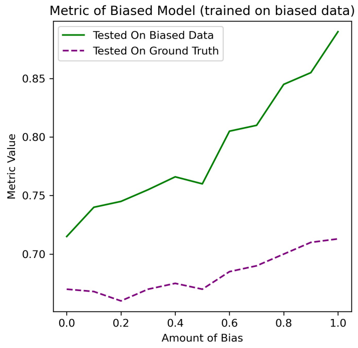


Figure 4.2: Exemplary visualization generated by the sandbox. We compare the performance of a biased model on ground truth and biased data.

under-representation bias into the training data and utilize Fairlearn's `CorrelationRemover` pre-processing algorithm to modify the data by removing correlations between the sensitive feature and the other features before training the model. What we observe is that, if we only evaluate on biased data, then we might be lulled into a false sense of progress and claim that the intervention is improving our model for increasing amounts of bias. However, when we examine the model's performance on the unbiased ground truth data, we see that performance does not improve significantly.

Overall, the sandbox tool regards the initial data set as unbiased and splits into training and test examples. While the training data are injected with bias, data reserved for testing remains untouched. After model fitting and fairness intervention, evaluation metrics and visualizations are provided on both the biased training data and the ground truth test data. The entire process is repeated for different levels of injected bias and, if indicated by the user, for several repetitions in order to obtain reliable average results.

## 4.3 Case study: can fairness constraints improve accuracy?

A main objective of the proposed sandbox tool is to aid empirical evaluation of the performance of fairness intervention under different biases. There are various special cases in which the effect of imposing fairness constraints has been characterized from a theoretical perspective (Khani and Liang, 2021; Zhou et al., 2021; Du and Wu, 2021). However, results like these usually focus on an infinite data setting and require a vast array of assumptions which can call their practical usefulness into question. In the coming sections, we use our sandbox tool to empirically replicate a known result from Blum and Stangl (2020) (Section 4.3) and explore performance beyond the assumptions required for the theory (Section 4.4). On a high level, we find that an often prohibitive amount of data is required to approximate the infinite data level result. In addition, our exploration suggests that the theoretical result breaks down completely if some of the structural assumptions on the problem setup and bias type are relaxed. The case study demonstrates how our sandbox tool can facilitate understanding of empirical implications of theoretical results and give the user a better sense of what performance to expect in their specific setting.[1]

We note that the case study and explorations discussed in the main text make various simplifying assumptions including an absence of confounding. Appendix C.3 presents supplementary experiments with confounding bias in a more realistic data setting.

### 4.3.1 Under-representation bias under Equalized Odds constraints

Fairness intervention into machine learning systems is often cast in terms of a fairness-accuracy trade-off. Yet learning from biased data can actually lead to sub-optimal accuracy once evaluated with regards to the unbiased data distribution. Blum and Stangl (2020) theoretically describe settings in which fairness-constrained optimization on biased data recovers the Bayes optimal classifier on the true data distribution. In this case study, we specifically zoom into one of the findings of the paper, that is, Equalized Odds constrained empirical risk minimization on data with under-representation bias can recover the Bayes optimal classifier on the unbiased data. This result requires several structural assumptions on the data generating process as outlined below. We will draw on the described data generating procedure when simulating data for the sandbox demonstration.

**Data generating process**. Let $G \in \{A, B\}$ specify membership in demographic groups $A, B$ where $B$ is the minority group and let $\mathbf{x} \in \mathcal{X}$ be a feature vector from some feature space. We assume there is a coordinate in $\mathbf{x}$ which corresponds to the group membership and write $\mathbf{x} \in A$ if individual $\mathbf{x}$ belongs to group $A$. The respective features distributions are denoted by $\mathcal{D}_A$ and $\mathcal{D}_B$. In order to generate data, we start with a pair of Bayes optimal classifiers $h^* = (h^*_A, h^*_B) \in \mathcal{H} \times \mathcal{H}$ where $\mathcal{H} = \{h : \mathcal{X} \to \{0, 1\}\}$ is a hypothesis class. For a given constant $r \in (0, 0.5]$, we then draw data points $x$ such that with probability

[1] The code generating the results in this Section can be found in the following repository: https://anonymous.4open.science/r/bias-stress-test-sandbox

$1 - r$ it holds $\mathbf{x} \sim \mathcal{D}_A$ and with probability $r$ it holds $\mathbf{x} \sim \mathcal{D}_B$. Dependent on the class membership, the true label is generated by first, using $h^*_A$ or $h^*_B$ and second, independently flipping the output with probability $\eta < 0.5$. The second step controls the errors of $h^*$ by ensuring that $h^*_A$ and $h^*_B$ have the same error rate and errors are uniformly distributed.

Starting with a ground truth data set including $m$ observations and label noise $\eta$, under-representation bias is introduced by discarding positive observations from the minority group $B$ with some probability. Specifically, for each pair $(\mathbf{x}, y)$ with $\mathbf{x} \in B$ and $y = 1$, the data point is independently excluded from the data set with probability $1 - \beta$. Note that $(1 - \beta)$ is the amount of under-representation bias.

**Recovery of the Bayes optimal classifier**. We first note that recovery of a classifier only pertains to the binary predictions. A Bayes optimal classifier learned from the noisy unbiased data does not necessarily have the same class probability predictions as $h^*$ even in the infinite data setting. To see this, consider the case in which $P(h^*(\mathbf{x}) = 1|\mathbf{x} \in A) = P(h^*(\mathbf{x}) = 1|\mathbf{x} \in B) = 1$ and $\eta = 0.2$. Then, fitting a sufficiently complex threshold based classifier on enough noisy data will result in a predictor $\hat{h}$ with $P(\hat{h}(\mathbf{x}) = 1|\mathbf{x} \in A) = P(\hat{h}(\mathbf{x}) = 1|\mathbf{x} \in B) = 0.8$. While class probabilities differ, both $h^*$ and $\hat{h}$ are Bayes optimal and, in this case, reflect the same binary predictor when selecting a threshold smaller or equal to 0.8.

**Main recovery result**. The derivations in Blum and Stangl (2020) are concerned with fairness constrained empirical risk minimization where an estimator $\hat{Y}$ is deemed fair if $\hat{Y} \perp G | Y = y$ for $y = 1$ (equality of opportunity) or $y \in \{0, 1\}$ (Equalized Odds). Here, $G$ denotes the protected group attribute. In our binary prediction setting, the Equalized Odds (Hardt et al., 2016) constraint is equivalent to

$$P\left(\hat{Y} \middle| \mathbf{x} \in A, Y = y\right) = P\left(\hat{Y} \middle| \mathbf{x} \in B, Y = y\right),$$

for $y \in \{0, 1\}$. The main result presented here is based on Theorem 4.1 in Blum and Stangl (2020) where a proof can be found. We note that this is a population level or 'with enough data' type of result.

**Theorem 13** (Blum and Stangl (2020))**.** *Let true labels be generated by the described data generating process and corrupted with under-representation bias. Assume that*

1. *both groups have the same base rates, i.e.* $p = P(h^*_A(\mathbf{x}) = 1|\mathbf{x} \in A) = P(h^*_B(\mathbf{x}) = 1|\mathbf{x} \in B)$*, and*

2. *label noise $\eta \in [0, 0.5)$ and bias parameter $\beta \in (0, 1]$ are such that*

$$(1 - r)(1 - 2\eta) + r(1 - \eta)\beta > 0.$$

*Then, $h^* = (h^*_A, h^*_B)$ is among the classifiers with lowest error on the biased data that satisfies Equalized Odds.*

### 4.3.2 Empirical replication using the sandbox toolkit

**Contribution of the sandbox**. The finding in Theorem 13 implies that fairness intervention can improve accuracy in some settings which goes against the common framing of fairness and accuracy as a trade-off. However, Theorem 13 is a purely theoretical result which can make it difficult to assess its usefulness in any specific application setting. For example, the Theorem operates at the population level suppressing issues of sample complexity. In practice it is unclear how much data would be needed for a satisfactory performance even if all the assumptions were met. Our proposed sandbox tool can bridge this gap between theory and practice by providing a controlled environment to test the effectiveness of fairness interventions in different settings. In the case of Theorem 13, the fairness sandbox can help to (1) Give a sense of how fast the result kicks in with a finite sample, (2) Assess effectiveness in a specific data generation and hypothesis class setting , and (3) Understand the importance of the different assumptions for the result.

**Implementation with the sandbox**. We describe how the different modules of the sandbox toolkit are used to empirically replicate the findings of Theorem 13.

1. **Choice of Data**: We opt for a synthetic data set generation according to the exact process described in Section 4.3.1. This leaves room for several input parameters which can be varied by the user. While some of these parameters determine whether the assumptions of Theorem 13 are met, i.e. the relative size of groups and the amount of label noise, the theorem is agnostic to the number of features, the distribution of features, and the Bayes optimal classifiers and their hypothesis class. In order to simplify reasoning about the results, our analysis focuses on a setting with only three features $x_1, x_2, x_3 \sim \mathcal{N}(0, 1)$ and a linear function class for the Bayes optimal classifiers. The illustration can be readily repeated for more features and a different Bayes Optimal classifier. But this simple example suffices to illustrate some of the key limitations of the theory in Blum and Stangl

(2020). More specifically, the group dependent Bayes optimal classifiers $h_A^*, h_B^*$ are thresholded versions of logistic regression functions

$$\log \frac{p}{1-p} = b_1 x_1 + b_2 x_2 + b_3 x_3 \tag{4.1}$$

for group dependent parameter vectors $\mathbf{b} \in \{\mathbf{b_A}^*, \mathbf{b_B}^*\}$. We set the parameters to fixed values $\mathbf{b_A}^* = (-0.7, 0.5, 1.5)^T$ and $\mathbf{b_B}^* = (0.5, -0.2, 0.1)^T$ which leads to different continuous distributions of probabilities between groups but to approximately the same positive rates when thresholded at 0.5 as required for the theoretical setting of the Theorem. Note that to adhere to the theory, we start out with a threshold-based classifier and subsequently add label noise with $\eta$ (instead of the more common way of turning probabilistic predictions into labels, i.e., flipping biased coins for the binary labels).

2. **Bias Injection:** Theorem 13 is concerned with a specific form of inter-sectional under-representation bias which strategically leaves out positive observations from the minority group. The sandbox is set up to inject this type of bias based on a user specified parameter $\beta$ which determines the amount of bias injected. The addition of further types of biases goes beyond the theory presented in Blum and Stangl (2020) and is empirically explored with the sandbox tool in Section 4.4.

3. **Model Class Selection:** The theoretical result we are looking to replicate operates on a population level and does not constrain the Bayes optimal classifier or learned model to belong to a specific class of functions. However, in practice we need to select a class of models with enough capacity to express both Bayes optimal classifiers $h_A^*$ and $h_B^*$ at once since the fairness constrained empirical risk minimization requires us to train a single model for both groups. To accomplish this, we select a logistic regression function of the form

$$\begin{aligned} \log \frac{p}{1-p} &= b_0 + \mathbf{1}(\mathbf{x} \in A)\mathbf{b_A}^T\mathbf{x} + \mathbf{1}(\mathbf{x} \in B)\mathbf{b_B}^T\mathbf{x} \\ &= b_0 + \begin{bmatrix} \mathbf{b_A} \\ \mathbf{b_B} \end{bmatrix}^T \mathbf{x}', \end{aligned} \tag{4.2}$$

where $\mathbf{b_A}$ corresponds to the parameters used for rows belonging to group $A$, and $\mathbf{b_B}$ denotes the parameters used for $\mathbf{x} \in B$. The indicator functions are absorbed into the data by reformatting the feature vectors $\mathbf{x} \in \mathbb{R}^3$ to feature vectors $\mathbf{x}' \in \mathbb{R}^6$ with $\mathbf{x}'^T = [\mathbf{x}^T, 0, 0, 0]$ for $\mathbf{x} \in A$ and $\mathbf{x}'^T = [0, 0, 0, x^T]$ for $\mathbf{x} \in B$. Note that the additional intercept $b_0$ increases the capacity of the model and can only help our performance here.

4. **Fairness Intervention:** Recall that Blum and Stangl (2020) analyze the setting of fairness constrained empirical risk minimization. We choose Equalized Odds constrained optimization as fairness intervention in order to mimic the theoretical setting of the result we are replicating. The constrained optimization is performed by scikit-learn unpenalized logistic regression with Equalized Odds enforcement provided by Fairlearn's Grid Search function which is based on Agarwal et al. (2018a). For the sake of comparison, we also fit the model from Equation 4.2 without fairness intervention.

   Since in-processing fairness intervention is not always desirable or possible, e.g. sometimes we only have access to biased black-box predictions, we conduct the same experiments with Fairlearn's post-processing method which enforces Equalized Odds by optimizing group-specific thresholds (Hardt et al., 2016). The respective results are discussed in detail in Appendices C.2.0.1 and C.2.0.2.

5. **Evaluation Metrics:** There are several relevant evaluation metrics for the case study, all of which are supported by our sandbox toolkit. First, we are interested in the overall and group-wise accuracy of the learned model which is provided for the models learned with and without fairness invention. Second, we evaluate the Equalized Odds disparity of the models in order to demonstrate the effectiveness of the intervention. Following Agarwal et al. (2018a), the extent to which a classifier $\hat{f}$ violates Equalized Odds is computed by

   $$\text{disp}(\hat{f}) = \max_{g,y} |\mathbb{E}[\hat{f}(\mathbf{x})|G = g, Y = y] - \mathbb{E}[\hat{f}(\mathbf{x})|Y = y]|,$$

   where $G$ is the protected group attribute. This definition is adapted to a finite data version by inserting the respective sample means for the expected values. Lastly, we want to demonstrate the explicit finding of Theorem 13 which is concerned with the recovery of Bayes optimal classifier. To this end, we compute the fidelity between the predictions of the learned models and the Bayes optimal classifier. The fidelity between two binary classifiers $\hat{f}_1$ and $\hat{f}_2$ with respect to a data set $D$

is defined as

$$\mathrm{fid}_D(\hat{f}_1, \hat{f}_2) = \frac{1}{|D|} \sum_{x \in D} \left| \hat{f}_1(\mathbf{x}) - \hat{f}_2(\mathbf{x}) \right| ,$$

i.e. as the fraction of examples on which the predictions of the classifiers coincide. The evaluation metrics are output each for the training and test sets. While fidelity results are discussed in detail in the main text, we refer to Appendix C.1 for a summary of accuracy and disparity results.

6. **Visualization:** The sandbox tool provides visualizations of the effectiveness of the fairness intervention. In the context of the case study, this consists of figures displaying the accuracies and fidelities to the Bayes optimal classifier of the models learned with and without fairness intervention at different levels of injected inter-sectional under-representation bias.

### 4.3.3 Empirical results

**Parameter inputs**. The sandbox tool with the described configurations is used to examine the empirical performance of the theoretical result from Blum and Stangl (2020) presented in Theorem 13. In this setting of the sandbox, the user can input several numerical values corresponding to the size of the minority group $r \in (0, 0.5]$, the number of synthetic data points to be generated $n \in \mathbb{N}$, the amount of overall label noise to be injected $\eta \in [0, 0.5)$ and the number of times the whole simulation should be repeated $R$. In each run of the simulation, new data are sampled and injected with bias before the respective models are fit. The whole simulation pipeline is performed based on the input values and performance metrics and visualizations are output to the user.

For the sake of demonstration, we chose $r = 0.2$, $\eta = 0.4$, $R = 50$, which provides one of many examples within the bounds of the theory. In an effort to explore how much data are actually required to obtain the performance promised by the population level theory in our example, we vary the number samples $n \in \{600, 6000, 60000\}$. We note that half of the synthetically generated data are used for model training and half for evaluation and visualization.

**Results**. Figure 4.3 displays the fidelity results of the sandbox simulation case study measured on the portion of the data sets withheld for testing. We note that fidelity here corresponds to the fraction of test examples that receive the same predictions from the model trained on biased data and the Bayes optimal classifier fit to the unbiased data. No bias is injected in the data used for testing.

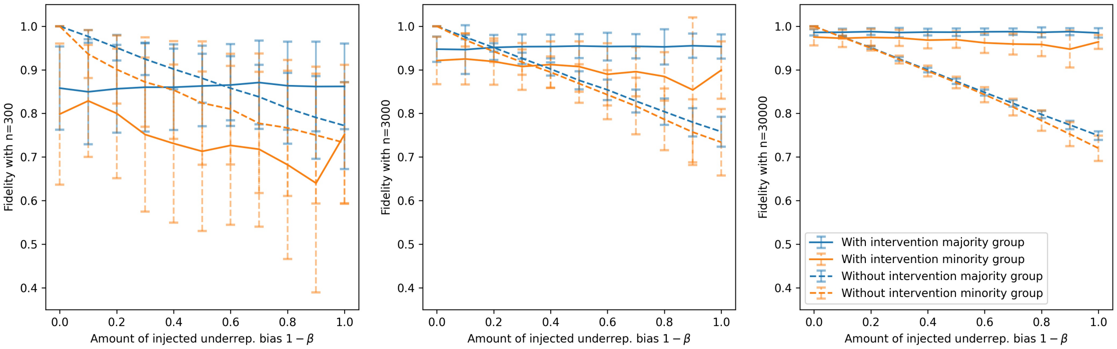


Figure 4.3: Test set fidelity between Bayes optimal classifier and models trained on biased data with and without fairness intervention using $n = 300, 3000, 30000$ (left to right) samples for training and testing each. Results are reported averaged over 50 simulation runs with error bars for one standard deviation in each direction. We see that Equalized Odds constrained optimization retrieves the Bayes optimal classifier almost perfectly at all levels of bias when using large amounts of data ($n = 30000$) but deviates from the Bayes optimal predictions when trained on $n \in \{300, 3000\}$ data points. The model class used in this example is logistic regression in 7 parameters.

We intuitively expect the model fit on biased data without fairness intervention to deviate from the Bayes optimal model especially when large amounts of bias are injected. This is confirmed by the downward slopes of the dashed curves in Figure 4.3. Theorem 13 implies that fitting the same model on biased data with Equalized Odds fairness intervention recovers the Bayes optimal classifier on the true data distribution. To see that the assumptions of the Theorem are met in our example, note that we selected the Bayes optimal classifiers $h^*_A$ and $h^*_B$ specifically to have equal base rates (see Equation 4.1), and that our choice of parameters $r = 0.2$ and $\eta = 0.4$ fulfills $(1-r)(1-2\eta) + r(1-\eta)\beta > 0$ for all levels of injected bias $1-\beta \in [0, 1)$. We would thus expect the fidelity of the models with fairness intervention to be 1 for all levels of $1-\beta$ which is only partially supported by Figure 4.3. For small amounts of training data ($n = 300$), the average fidelity over simulation runs and levels of injected bias only reaches a level of 0.837 with even poorer performance in the minority group. In cases with 90% of positive minority examples deleted from the training data, the model learned with fairness intervention on average only classifies about 64% of the minority test examples the same way as the Bayes optimal classifier. In addition, results vary significantly over simulation runs leading to many instances with little to moderate amounts of injected bias in which the model learned from biased data without intervention is closer to the Bayes optimal than the model with intervention. With more training data ($n = 3000$), the test fidelity performance of the intervention model increases to 0.942 on average. Yet even in this setting, the biased model outperforms the intervention model if only 20% or less of positive minority examples are deleted from the test data. Only when increasing the training data size to ($n = 30000$), the fidelity of the intervention model reaches 0.982 which is much closer to the results implied by the theory. In this case, the model with intervention outperforms the model without intervention for almost all positive bias levels.

Overall, the findings of the sandbox demonstrate that a considerable amount of data are needed to recover from under-representation bias. We only observed satisfactory results at all positive bias values when 30000 training examples were used for a relatively simple 7 parameter logistic regression model. [2] Many practical applications fall into the range of small data sets and little to moderate under-representation bias in which the intervention model showed to be no more successful in recovering the Bayes optimal model than a model without intervention. The presented case study demonstrates how the sandbox toolkit can help to uncover insights of this type for users who are looking to assess the effectiveness of fairness intervention in their specific application setting.

**Comparison to post-processing intervention**. While Blum and Stangl (2020) specifically call for in-processing intervention, fairness constrained risk minimization is not the only method that targets Equalized Odds across groups. Since post-processing strategies are desirable in some cases, we repeat the same experiments with the threshold based post-processing Equalized Odds algorithm from Hardt et al. (2016). Note that this corresponds to changing the configuration of step '(4) Fairness intervention' in the sandbox pipeline while keeping the fairness metric fixed. The results from this analysis are discussed in Appendix C.2.0.1 and indicate a very similar performance to the in-processing method.

## 4.4 Exploration of other forms of bias

Section 4.3 demonstrates the usefulness of the proposed sandbox tool by empirically evaluating the performance of a theoretical result from Blum and Stangl (2020). For this, we assume the exact setting of the paper with requires a list of structural assumptions on the synthetic data generation, Bayes optimal model and type of injected bias. For example, the replicated finding only considers a specific case of under-representation bias. Real world applications are likely to violate some of the posed assumptions

[2] In general, the amount of data required to reliably fit a model increases with complexity of the model class. Many algorithms used in practice exceed the complexity of the model studied here which suggests that even more data are required to observe the desired fairness mitigation effects.

and can carry a number of different biases. In the following, we show how the modularity of the sandbox allows us to explore the performance of fairness intervention beyond the setting posed by the theory. We loosen the assumption of equal base rates in Bayes optimal predictions and inject different types of biases in order to stress-test the efficacy of the intervention. The changes to the sandbox modules discussed in the following refer to the sandbox configuration presented in Section 4.3.2.

### 4.4.1 Difference in base rates

**Implementation with the sandbox and parameter values**. The result of Theorem 13 relies on the assumption that base rates are the same across groups which is often violated in practice. We use the sandbox framework to test the extent to which the fidelity of the Equalized Odds intervention is affected by diverging rates and alter the data choice module of the sandbox used in the case study for this purpose. A collection of data sets with different base rates is generated as follows. We leave the labeling model and effect parameters $b^*_A, b^*_B$ untouched and sample the features $x_1, x_2, x_3$ conditional on group membership with $x_i|(\mathbf{x} \in A) \sim \mathcal{N}(d, 1)$ and $x_i|(\mathbf{x} \in B) \sim \mathcal{N}(0, 1)$ for $i = 1, 2, 3$. Here, $d$ is from a collection of feature mean values selected to lead to evenly spaced base rate differences in $[-0.5, 0.5]$ once the binary Bayes optimal outcomes are computed. Note that the rate of positive outcomes for the minority group $B$ is always 0.5 which justifies the range of the interval.

As in the previous experiments, we set the additional input parameters to $r = 0.2, \eta = 0.4$ and $R = 50$. This aligns with the setting in Section 4.3 and thus enables us to compare performance across different types of injected bias. That the choices made here are one example among many, they were picked early on to comply with theory and were never changed to obtain specific results. We run the experiment with $n = 60000$ data points at each base rate difference split evenly between training and testing and set the under-representation bias level to $1 - \beta = 0.4$.

**Results**. Figure 4.4 depicts the test set fidelity of the classifiers trained on biased data with Equalized Odds intervention and the data-driven Bayes optimal model at different levels of base rate difference between groups. The base rate difference is here defined as the base rate of the majority group minus the base rate of the minority group where latter is fixed at 0.5. While the rate of positive Bayes optimal outcomes in the minority group is constant at 0.5, the base rate in the majority group varies between 0 and 1 in our experiment. We see that the intervention model is able to recover the Bayes optimal classifier

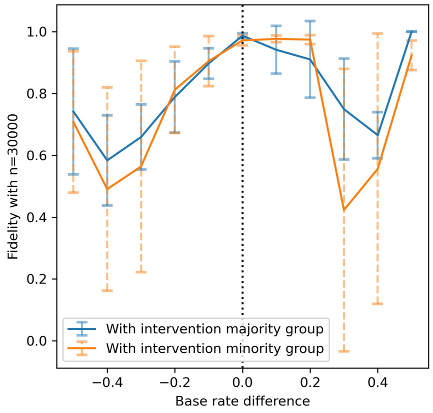


Figure 4.4: Test set fidelity between Bayes optimal classifier and model trained on biased data with Equalized Odds intervention. Results are reported as an average over 50 simulation runs. Error bars correspond to one standard deviation in each direction. We see that the fidelity between models is generally smaller than 1 if base rates are not the same across groups. In other words, the intervention fails to retrieve the Bayes optimal classifier in these cases.

for a base rate difference of 0 which corresponds exactly to the setting of Theorem 13. The larger the base rate difference becomes in absolute value, the more the predictions of the fair trained model and the Bayes optimal model diverge. The performance in the minority group appears to be particularly poor with a minority base rate of 0.5 and majority base rate of 0.8 leading to minority group fidelity of 0.423 on average. Larger differences in base rates also seem to lead to intervention models with less stable performance which leads to large standard errors.

In order to understand why the result of Theorem 13 does not generalize to settings with different base rates $p_A \neq p_B$, consider that the true positive rate of the Bayes optimal classifier for $G \in \{A, B\}$ on unbiased data takes the form

$$P(h^*_G(\mathbf{x}) = 1 | Y = 1, \mathbf{x} \in G) = \frac{(1-\eta)p_G}{p_G(1-\eta) + (1-p_G)\eta},$$

which is different for different base rates $p_A$ and $p_B$. When under-representation bias $1 - \beta \neq 1$ is introduced, the true positive rate for group $B$ becomes

$$P(h^*_B(\mathbf{x}) = 1 | Y = 1, \mathbf{x} \in B) = \frac{(1-\eta)\beta p_B}{p_B\beta(1-\eta) + (1-p_B)\beta\eta},$$

which coincides with the rate for the unbiased data. It follows that the Bayes optimal classifier does not

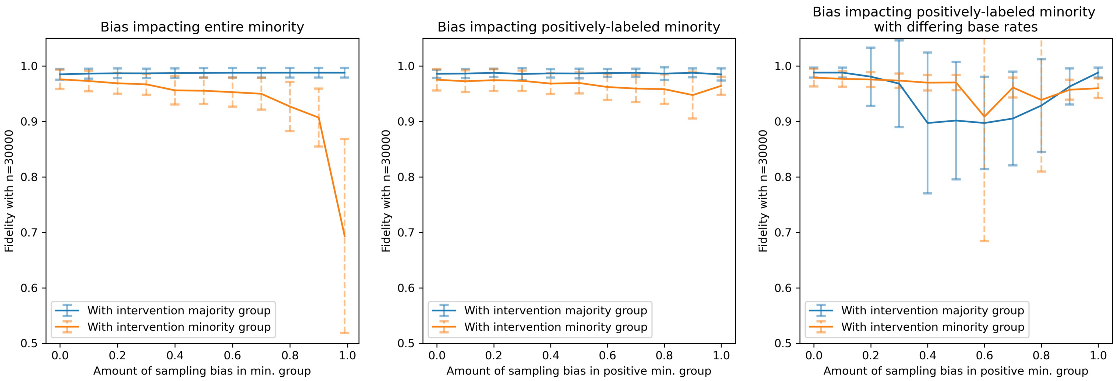


Figure 4.5: **Sampling bias.** Test set fidelity between Bayes optimal classifier and models trained on biased data with Equalized Odds intervention on 30000 samples for training and testing each. Results are reported averaged over 50 simulation runs with error bars for one standard deviation in each direction. Bias is injected into either the entire minority group (left), or the positively labeled minority group (middle). On the right, bias is injected into the positively labeled minority group and we assume a base rate difference of -0.2

have equal true positive rates, and thus does not satisfy Equalized Odds, on the biased data if base rates are different. It can therefore not be recovered by the fair trained model.

### 4.4.2 Sampling bias

**Implementation with the sandbox and parameter values**. Our previous discussion of under-representation bias only considered bias specifically injected into the subgroup of examples at the intersection of minority group and positive labels. We extend this setting to under-representation bias in the full minority group, which we will refer to as sampling bias, by altering the bias injection module of the sandbox to remove minority examples with some probability ranging between 0 and 1. Experiments are repeated with equal base rates and with base rate difference of -0.2 which allows us to explore how the performance changes as a difference in base rates is introduced while ensuring that the data still contains examples for both outcomes in each group. We set the parameter inputs to $r = 0.2, \eta = 0.4, R = 50$ and $n = 60000$ to comply with the parameter choices in previous experiments.

**Results**. The results of the experiments for bias injected in the whole minority group, positively labeled minority examples, and positively labeled minority examples with different base rates are depicted in the first column of Figure 4.5. The left plot shows a decreasing minority test set fidelity with increasing sampling bias in the minority group. With maximally injected bias, 99% of minority examples are deleted and the average minority group fidelity only reaches 0.694. With smaller amounts of bias, the intervention model classifies over 90% of minority test samples like the Bayes optimal classifier. Intuitively, the decreased performance on the minority set can be led back to less available training data for the group. Since we fit only one model for both groups, this leads the predictions for the majority group to be closer to the Bayes optimal predictions than for the minority group. When bias is injected only for positively labeled minority examples, the intervention successfully recovers the base optimal classifier as discussed in Section 4.3. The right plot of the figure displays the test set fidelity in the case of different Bayes optimal base rates in groups with bias injected only for positively labeled minority group examples. We note that the fidelity here appears much less stable over different runs of the simulation which leads to larger standard errors. In contrast to the setting with equal base rates, the bias injection here impacts also the fidelity of the majority group. Recall that the Bayes optimal classifier does not satisfy Equalized Odds on the biased data in this setting and can thus not be recovered by strictly requiring Equalized Odds. However, the figure suggests a remarkably high fidelity for low amounts of bias in the different base rates case and we hypothesize that the model was faced with large accuracy fairness trade-offs and opted for a small violation of the fairness constraint in favor of accuracy.

### 4.4.3 Label bias

**Implementation with the sandbox and parameter values**. Recall that our experiments use a noise parameter $\eta$ which represents the probability with which the Bayes optimal label is flipped in our observed labels. So far, this value was chosen independently from group membership. Since data in real-world application often suffers from differential label noise, we test how well the Equalized Odds intervention can recover the Bayes optimal model under label bias. To achieve this, we alter the choice of data module to inject 40% label noise into the majority group to be consistent with the previous experiments. We then change the bias injection module to inject label bias of 0-45% into the minority group. Note that the label bias cannot exceed 50% in order for the Bayes optimal classifiers to be correct which justifies the chosen range. Similarly, we repeat the experiment by injecting constant bias of 40% into both the majority and negatively labeled minority group and vary the amount of bias among the positively labeled minority. In all instances, the test set has 40% label bias throughout like in the previous experiments. The experiment is repeated with different base rates for bias injected into the positively labeled minority examples. As

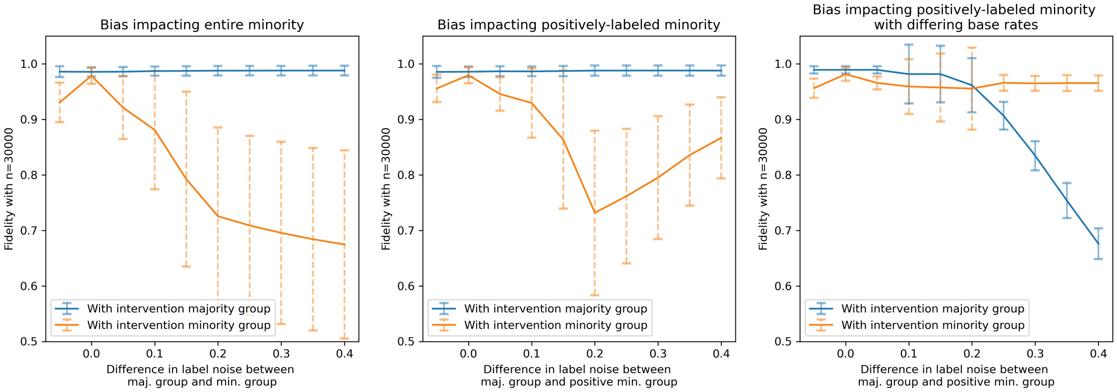


Figure 4.6: **Label bias.** Test set fidelity between Bayes optimal classifier and models trained on biased data with Equalized Odds intervention on 30000 samples for training and testing each. Results are reported averaged over 50 simulation runs with error bars for one standard deviation in each direction. Bias is injected into either the entire minority group (left), or the positively labeled minority group (middle). On the right, bias is injected into the positively labeled minority group and we assume a base rate difference of -0.2.

before, we set $r = 0.2$, $R = 50$ and $n = 60000$.

**Results**. The results of the label noise experiments are depicted in Figure 4.6. We observe that fidelity performance of the intervention model deteriorates in the minority group as the amount of label noise diverges. This holds true both when bias is injected into the whole group and when bias is injected into positively labeled minority examples. To understand why the intervention cannot retrieve the Bayes optimal classifier, we note that the Bayes optimal classifier does not fulfill Equalized Odds under differential label noise. To see this, assume a setting with $\eta_{\text{maj}} = 0.4$ label noise bias in the majority group and $\eta_{\text{min}} \neq 0.4$ bias in the minority group. The Bayes optimal classifier $h_A^*$ has true positive and true negative rates of 0.6 on the majority group data while $h_B^*$ has true positive and true negative rates of $1 - \eta_{\text{min}} \neq 0.6$ on the minority portion of the biased data. Note that this assumes that base rates are 0.5 like in our experiment, but the same phenomenon with a similar calculation holds true for other cases. If bias is injected into the positively labeled minority examples and base rates differ by -0.2, the fidelity curves of both the minority and majority groups are impacted.

### 4.4.4 Feature measurement bias

**Implementation with the sandbox and parameter values**. Our final exploration focuses on a type of feature noise which is injected in the form of missingness in one of the features. We alter the bias injection

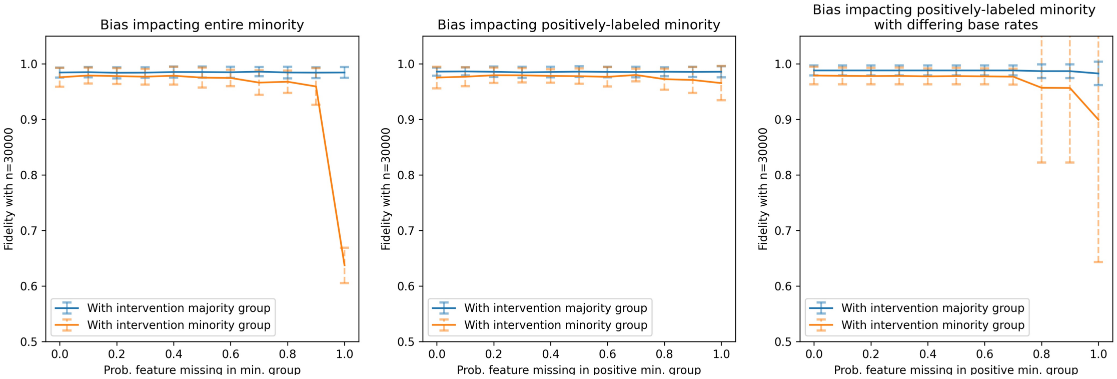


Figure 4.7: **Feature measurement bias.** Test set fidelity between Bayes optimal classifier and models trained on biased data with Equalized Odds intervention on 30000 samples for training and testing each. Results are reported averaged over 50 simulation runs with error bars for one standard deviation in each direction. Bias is injected into either the entire minority group (left), or the positively labeled minority group (middle). On the right, bias is injected into the positively labeled minority group and we assume a base rate difference of -0.2.

module in the sandbox and set feature $x_1$ to 0 with varying probability while omitting the injection of other types of biases. The functionality to enforce different base rates in the data choice module is retained. Feature measurement bias is injected in to the whole minority group or the minority group with positive labels in different variations of the experiment. As before, we choose $r = 0.2, \eta = 0.4, R = 50$ and $n = 60000$. Experiments are repeated with base rate differences of 0 and -0.2.

**Results**. The test set fidelity results of the feature noise experiments are displayed in Figure 4.7. We see that the intervention model recovers the Bayes optimal model for small amounts of feature missingness while fidelity slightly decreases as more bias is injected. While performance remains above the 0.95 mark for most amounts of injected bias, we observe an average fidelity of only 0.635 if all instances of $x_1$ in the minority group default to 0. Similar to the other types of bias, the intervention model successfully recovers from measurement bias if the bias is only injected into positively labeled examples. In contrast to our observations for other types of biases, a difference in base rates appears to not deteriorate the fidelity by more than 1-2 percentage points for up to 70% feature missingness in a single feature among the minority group examples with positive labels. Assuming our hypothesis from Section 4.4.2, i.e. with no additional bias the fair learned model on different base rates trades off fairness for accuracy, is true, we conjecture that the performance when injecting measurement bias does not deteriorate quickly because it only introduces very small amounts of Equalized Odds disparity. On a high level, removing the information of one feature leads to a higher concentration around the mean in the predicted conditional probabilities. While this leads to a small violation of Equalized Odds, the fair trained model accepts this unfairness in favor of a high accuracy.

### 4.4.5 Comparison to post-processing intervention

We repeat the exploration experiments with threshold-based post-processing fairness intervention (Hardt et al., 2016) which corresponds to altering the fairness intervention module of the sandbox tool. The result are discussed in detail in Appendix C.2.0.2. While the two intervention methods showed to lead to fairly similar results in the original case study setting, this is not necessarily the case when base rates differ or different types of bias are injected. In those cases, the algorithms face a trade-odd between fairness and accuracy which can lead to different predictions across different intervention methods. For example, we see that the in-processing method yields higher fidelity performance than the post-processing intervention when feature measurement bias is injected as the in-processing method is better in trading off some amount of fairness for accuracy.

## 4.5 Related work

**Types of fairness-enhancing algorithms**. At a high level, there are three classes of fairness-enhancing algorithms or fairness interventions: pre, post, and in-processing (Zhong, 2018). These algorithms are applied at different stages in the ML pipeline and can be accommodated by our sandbox toolkit. Pre-processing algorithms modify the data itself and remove the underlying biases which is best suited when training data are accessible and modifiable. Examples of pre-processing algorithms include optimized pre-processing (Calmon et al., 2017), disparate impact remover (Feldman et al., 2015) and reweighing (Kamiran and Calders, 2012). In-processing algorithms operate on the model directly, removing biases during the training process. Examples in this category include the Meta-Fair Classifier (Celis et al., 2019), adversarial debiasing (Zhang et al., 2018) and exponentiated gradient reduction (Agarwal et al., 2018a). Post-processing algorithms utilize the predictions and modify the model outputs directly. This approach is best suited when neither the data nor the models are accessible, as it only requires access to black-box predictions. Example algorithms include Equalized Odds post-processing (Hardt et al., 2016) and reject option classification (Kamiran et al., 2012). In this chapter, we demonstrate how our sandbox toolkit applies both in-processing and post-processing fairness interventions at the example of a result from (Blum and Stangl, 2020).

**Fairness toolkits**. In order to ease application of fairness interventions in practice, recent work has developed a number of open-source ML fairness software packages or "fairness toolkits" (Bird et al., 2020; Bellamy et al., 2019; Saleiro et al., 2018; Bantilan, 2018; Wexler et al., 2019; Adebayo et al., 2016; Tramer et al., 2017). For example, Fairlearn (Bird et al., 2020) consists of an API to allow researchers and developers to easily use popular fairness interventions (such as Equalized Odds or Demographic Parity) at the three stages of the ML pipeline listed above. Most of these toolkits focus on *fairness interventions*, or how to apply fairness algorithms (Bird et al., 2020; Bellamy et al., 2019; Saleiro et al., 2018; Bantilan, 2018; Wexler et al., 2019; Adebayo et al., 2016; Tramer et al., 2017). A key distinguishing feature of our work is that our toolkit focuses on specific *biases* themselves. Our toolkit allows users to inject biases into their data, uses algorithms from Fairlearn to apply fairness interventions, and then compares to the ground truth. Our toolkit currently uses Fairlearn to apply fairness interventions due to its popularity and ease-of-use. Though, in future development of the toolkit, we plan to add other fairness toolkits, such as AIF360 (Bellamy et al., 2019).

**Algorithms for specific sources of unfairness**. A motivating reason why we focus on injecting specific biases in our toolkit is to evaluate or empirically replicate work or which claims to address specific sources of bias or unfairness. For example, in this chapter, we primarily focus on representation bias, measurement bias, and label bias (Frénay and Verleysen, 2013; Mehrabi et al., 2021). See Mehrabi et al. (2021) or Suresh and Guttag (2021) for further detail on more sources of bias or unfairness. To address these sources of unfairness, some have proposed solutions beyond algorithms, such as creating a more representative dataset[3] addressing larger societal inequities. In our toolkit, however, we focus on interventions which can be implemented at the time of training a model, after the dataset has already been created and any broader conditions surrounding model deployment are fixed. Recent work has proposed attempts at algorithmic solutions to remedy specific sources of biases or unfairness. Here, we present examples of the kinds of papers which our toolkit would be able to evaluate. For example, in the context of medical diagnoses,

[3] See, for example, https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html in response to (Shankar et al., 2017).

there exists a significant discrepancy in the quality of an evaluation (consider this as the label) between different races (Obermeyer et al., 2019). Khani and Liang (2021) show that removing spurious features (e.g. sensitive attributes) can decrease accuracy due to the inductive bias of overparameterized models. Zhou et al. (2021) finds that oversampling underrepresented groups can not only mitigate algorithmic bias in systems that consistently predict a favorable outcome for a certain group, but improve overall accuracy by mitigating class imbalance within data that leads to a bias towards majority class. Du and Wu (2021) observes the impact of fairness-aware learning under sample-selection bias. Wang et al. (2021) considers label bias based on differential label noise. Wang et al. (2020) looks at whether fairness criteria can be satisfied when the protected group information is noisy, missing, or unreliable. While our toolkit is able to address many of the claims in these papers, we focus on applying Equalized Odds intervention to data sets injected with the biases listed above in this chapter.

## 4.6 Summary and future directions

This work presented the idea and first implementation of a simulation toolkit to investigate the fairness consequences of various forms of biases and identify effective remedies for each the performance of fairness-enhancing algorithms under various forms of counterfactually injected biases. We demonstrated the utility of our tool through a thorough case study of Blum and Stangl (2020). The theoretical contribution of Blum and Stangl (2020) stated that if the source of unfairness is under-representation bias in the training data, constraining ERM with EO can recover the Bayes optimal classifiers on the unbiased data under certain conditions. Our tool allowed us to examine EO constraints under the conditions of Blum and Stangl (2020) as well as a number of new biased settings.

**Lessons from case study**. Through our case study, we established several limitations of the existing theory. In particular, we observed the need for very large volumes of data for the theory to hold. (In our example, we needed 30k training data points for a 7-parameter logistic regression.) Furthermore, our empirical results suggest that the smaller the amount of injected bias, the larger the volume of data needed in order for the fairness-constrained model to outperform the unconstrained one trained on biased data. We emphasize that many practical applications do not satisfy these preconditions (i.e., either the volume of data or the amount of under-representation bias is relatively small). Therefore the theoretical findings of Blum and Stangl (2020), while conceptually interesting, might not be applicable in those practical domains. Another key prerequisite of the theory was the equality of base rates across groups. This assumption is also often violated in practice, and we showed empirically that EO constrained ERM can not recover the Bayes optimal models if base rates differ—even slightly.

**Exploring the implications of various biases and interventions**. We experimented with various forms of biases and assessed the performance of EO constraints in alleviating them. For example, our empirical investigation of *sampling bias* demonstrated how the EO-constrained model struggles to recover comparable performance across groups. We also observed that the constraint could not retrieve the Bayes optimal classifiers under *label bias* either. In terms of the choice of interventions, we contrasted the in-processing method of Agarwal et al. (2018a) with the post-processing method proposed by Hardt et al. (2016). The key distinction between these two approaches appeared to be in their ability to trade off accuracy and fairness. In particular, the in-processing method offers a wider range of tradeoff possibilities, while the post-processing method yields fair classifiers but with no error guarantees. When the theoretical conditions of our case study hold, the two methods perform similarly, but they diverge as soon as those conditions are relaxed.

**Scope of applicability and limitations**. First, we should emphasize that the sandbox tool should be understood as an environment to explore the limitations of various fairness interventions in user-specified biased settings rather than a method to obtain fully generalizable results. The insights obtained through this exploration can form the basis of informed hypotheses for further empirical and/or theoretical investigations, but on their own they do not guarantee generalizability. For example, the analysis presented in Section 4.4 reveals that, at least in our specific experimental setting, EO-constrained optimization cannot recover the Bayes optimal classifier when base rates between groups differ or the data are impacted by label bias. While these findings are not guaranteed to hold in settings beyond the ones studied here, they allow us to surface several limitations of EO constraints as fairness interventions.

Second, we note that the current version of our tool is designed with the intention of helping researchers and students to form a better understanding of sources of unfairness. Our implementation of the data biases mentioned in this work is highly simplified, and it does not capture the complex nature of bias in real-world data. Addressing bias in specific domains requires prolonged deliberations with domain-experts and stakeholders. Therefore, the results obtained using our tool should not be interpreted in vacuum as the *proof* of efficacy (or lack thereof) for a given algorithmic fairness interventions *in practice*.

**An active-learning module in AI ethics curricula**. In recent years, call for "greater integration of ethics across computer science curriculum" have amplified (see, e.g., Fiesler et al. (2020)). However, instructors without a background in the area may lack the necessary tools to cover these issues in depth (Saltz et al., 2019; Martin, 1997). Our sandbox toolkit can serve these educators as a self-contained and ready-to-use learning module. With the toolkit, students can inject various types of biases into a given dataset, observe the fairness ramifications of the bias, and evaluate the effectiveness of various fairness interventions in alleviating them. By offering a hands-on practice, we hypothesize that the toolkit improves students' understanding of the Machine Learning pipeline, the underlying causes of unfairness, and the scope and limitation of existing algorithmic remedies depending on the type of bias present in the setting at hand. In our future work, we plan to conduct human-subject studies at college-level computer science programs to examine the effect of our sandbox toolkit in achieving FATE-related learning objectives and improving the learning experience.

# Part II

# Dynamics of fairness-promoting interventions

Chapter 5

# Long-term dynamics of fairness intervention in connection recommender systems

Based on (Akpinar et al., 2022a): Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. Long-term dynamics of fairness intervention in connection recommender systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (AIES 2022).

## 5.1 Introduction

Machine learning based recommender systems are at the heart of user experience in many social media applications. These systems underpin a wide range of services, including content ranking, connection recommendation, and job search tools. It is imperative that people participating in these systems, either as the recipient of ranked suggestions, or as the originator of the content being recommended are treated fairly which motivates a rich body of research in ranking and recommendation fairness (e.g. Zehlike and Castillo, 2020; Yang and Stoyanovich, 2017; Singh and Joachims, 2019; Zehlike et al., 2017; Elisa Celis et al., 2018).

Consider a connection recommendation setting where the system suggests a list of users based on a prompt such as ‘People you may know’ and the recipient of the recommendation decides which of the users to connect with. How should the platform promote fairness between different user groups in these recommendations? An array of definitions, fairness enhancing algorithms and evaluation metrics have been proposed to address this and similar problems. Most approaches assume static prediction settings and focus on a single fairness metric in individual instances of the recommendation in a one-shot or time-aggregate manner. However, recommender systems are dynamic in nature with recommendations influencing user behavior and experience through time. This is particularly evident in people connection recommendation systems which lead to a connection graph that evolves over time. Limiting ourselves to only one targeted fairness metric while neglecting other important variables and potential dynamics may lead to unintended consequences and overlooked side effects (Dai et al., 2021; D'Amour et al., 2020).

This work focuses on the long-term effects of fairness intervention in connection recommendation. We empirically demonstrate that these systems can suffer from a group-wise 'rich-get-richer' phenomenon which exacerbates outcome disparities over time. Through a simulation framework, we study the long-term impact of fairness interventions, finding that, although seemingly fair in aggregate, a key desiderata of fairness intervention, i.e. equity in network sizes, is not promoted over time. In fact, average network sizes diverge in the long run even with popular fairness interventions leaving the bulk of the minority group disadvantaged while simultaneously creating an illusion of fairness. We support our empirical findings by conducting a theoretical limit analysis of the impact of different fairness interventions on bias amplification dynamics assuming a stylized connection recommendation system based on Pólya urns.

Studying the potential long-term harms of mitigation approaches through online experimentation can be time consuming and potentially lead to real harm. Also, because of issues of network interference, it can be challenging to fully understand the impacts of fairness interventions on a network graph through experimentation. For these reasons, we defer to simulation along with theory that supports the simulation findings. Simulation-driven methods to uncover long-term effects have been previously used in the recommendation literature following the observation that offline experiments on observational data are often insufficient to assess performance after deployment (Krauth et al., 2020; Gomez-Uribe and Hunt, 2016), and they provide a promising path towards understanding context specific fairness dynamics in the ranking and recommendation fairness setting (Patro et al., 2022).

The remainder of the chapter is organized as follows. Section 5.2 provides an overview of related fairness literature including fairness in recommender systems, mitigation approaches, and long-term dynamics. Section 5.3 outlines fairness criteria, corresponding methodology for mitigation, and a simulator modelling a connection recommender system. Results of the simulation study are given in Section 5.4. Section 5.5 derives theoretical results demonstrating the validity of the empirical results, and leverages this theory to understand the workings of the fairness interventions. Finally, we conclude with a discussion in Section 5.6.

## 5.2 Background and related work

### 5.2.1 Fairness in recommendations

The most well-studied types of recommender system bias include popularity bias which refers to the over-recommendation of already popular items (e.g. Jannach et al., 2015; Abdollahpouri and Mansoury, 2020; Abdollahpouri et al., 2019b), and position bias which describes the tendency of members to interact primarily with top ranked items when shown recommendations in the form of a list (e.g. Joachims et al., 2007; 2017; Craswell et al., 2008; Wang et al., 2018). More recently, researchers have started to take interest in outcome disparities on a group level which can have an intricate relationship with known deficiencies like popularity bias (Ekstrand et al., 2018; Abdollahpouri et al., 2019b). While some work considers group level disparities on the source side, i.e. for the members who implicitly query and receive the recommendations (e.g. Ekstrand et al., 2018), a considerable body of research concentrates on fairness for the destination side, i.e. the items or people being recommended. This focus is based on the common understanding that exposure in recommendations is a valuable but scarce resource that can be the deciding factor in which suppliers can sell their items or who gets a job offer.

A variety of different metrics and bias mitigation algorithms for fairness in recommendation lists have been proposed (e.g. Mehrotra et al., 2018; Zehlike et al., 2017; Yang and Stoyanovich, 2017; Elisa Celis et al., 2018; Beutel et al., 2019; Nandy et al., 2020). For example, (Mehrotra et al., 2018) use counterfactual estimation techniques to understand the impact of different recommendation policies in the context of music recommendation which has been shown to suffer from gender bias (Ferraro et al., 2021; Shakespeare et al., 2020). A different line of work relies on pairwise comparisons from randomized experiments to measure fairness in rankings (Beutel et al., 2019). The works of (Zehlike et al., 2017) and (Yang and Stoyanovich, 2017) are concerned with the problem of top-$k$ ranking, and the authors of (Nandy et al., 2020) propose post-processing methods to achieve equality of opportunity or equalized odds in recommender systems. A straightforward but flexible statistical notion of fairness in the context of recommendations is demographic parity of exposure which has been used by a number of papers. (Singh and Joachims, 2018) maximize ranking utility for the viewer of a ranked list subject to exposure-centered fairness constraints including demographic parity. The work of (Zehlike and Castillo, 2020) uses an in-processing approach that directly focuses on enforcing demographic parity of exposure in scoring models used for ranking. (Abdollahpouri et al., 2019a) analyze how popularity bias affects different stakeholders of recommender systems and propose exposure-based metrics such as demographic parity to measure unfairness.

Based on the popularity of the metric, our analysis of long-term dynamics of fairness interventions in connection recommender systems begins by assuming demographic parity of exposure as fairness measure. The empirical portion of this work assumes a probabilistic ranking framework similar to the settings in (Basu et al., 2020; Singh and Joachims, 2018), and tracks the effects of demographic parity of exposure and other fairness metrics in connection recommendation over an extended period of time which enables us to make observations that have previously been overlooked.

### 5.2.2 Feedback loops and long-term fairness

Evaluation of fairness in recommender systems often focuses on single recommendation steps or time-aggregate behavior of the system. An exception to this is a body of work on popularity bias which has been shown to be highly dynamic over time (Yao et al., 2021; Chen et al., 2020). In many applications, popularity bias leads to a feedback loop that further increases the exposure of popular items over time leading to a long-tail phenomenon often called “rich-get-richer effect” or “Matthew effect”. (Chen et al., 2020) summarize methods proposed to break this feedback loop including reliance on uniform data (Jiang et al., 2019; Liu et al., 2020) and reinforcement learning-based recommenders which are able to adapt to changing states of the system (Zhao et al., 2019; Ge et al., 2021). Both of these approaches are challenging to realize in real-world applications. While obtaining uniform data in practice generally requires deploying some sort of uniformly at random recommendation policy which hurts member experience (Chen et al., 2020), reinforcement learning-based recommenders are difficult to evaluate since they only have access to data biased by an existing policy (Jagerman et al., 2019; Chen et al., 2019). To the best of our knowledge, few works have considered how intervening on fairness between different demographic groups may change the dynamics of a recommender system in the long term. (Ferraro et al., 2021) study the setting of music recommendation and observe a positive feedback loop when intervening on gender fairness. The authors propose an iterative reranking approach to mitigate unfairness measured based on a number of task-specific metrics. Simulation results suggest that, over time, intervention leads the music recommendation algorithm to make fairer recommendations organically. (Morik et al., 2020) incorporate a potential feedback loop into their intervention procedure, and propose an algorithm in the form of a controller that optimizes utility under amortized group fairness constraints while dynamically adapting as more data becomes available.

In this work, we study how the addition of common statistical notions of fairness to probabilistic ranking problems for connection recommendation impacts the state of a social network in the long term. Our variable of interest is the difference in average network sizes between groups which can be understood as a measure of diverging benefits.This follows a similar idea as (Liu et al., 2018) who, while not directly concerned with recommender systems, study the delayed impact of fairness intervention in a classification setting with a one-step feedback loop. The authors assume a hypothetical lending scenario in which lending decisions are based on and impact the score distributions in two demographic groups. Here, the central quantity of interest is the difference between average scores in groups and the findings suggest that common fairness criteria do not necessarily promote improvement over time.

## 5.3 Methodology

### 5.3.1 Optimization framework for connection recommendation

Connection recommender systems generally rely on a member's current network as well as other member data in order to select a group of members of the platform to suggest as potential connections. Expanding one's network is assumed to create positive value for the both the member at the source side (viewer of the recommendation) and the member at the destination side (recommended member). Recommendations generally do not require specific search terms or prompts by the source member but are instead often provided in a single or a few automatically generated categories, e.g. 'People you may know' or 'People you may know from your workplace'. Each member of the platform simultaneously serves as a source and a destination member and is thus associated to a two dimensional vector of recommendation utilities. Despite stakeholders participating at both sides, connection recommender systems generally focus on source side utility similar to other types of recommender systems (Agarwal and Chen, 2016; Liu, 2007; Adomavicius and Tuzhilin, 2005). We build on the framework of (Singh and Joachims, 2018; Basu et al., 2020) and formalize the recommendation problem as follows.

Let $s$ denote a source member (initiating a query to the recommender system) and $d$ a destination member (candidate to be shown as a recommendation), and let the implicit query for recommendations $q$ yield an ordered list of $m$ destination members. The relevance or ranking score for the pair $(s, d)$ is denoted by $u^q_{s,d}$ and reflects an abstract quantity of utility that member $s$ receives from being recommended member $d$. We assume a ranking policy matrix $P^q_s \in \mathbb{R}^{D_q \times m}$ where $D_q \geq m$ is the total number of eligible destination members for the query and $P^q_s(d, r)$ denotes the probability with which member $d$ is shown to member $s$ in slot $r$ of the recommendation list. Lastly, $v \in \mathbb{R}^m$ is a fixed vector that models position bias by encoding how much attention destination members pay to recommendations in slots $r = 1, \ldots, m$. This exposure vector is generally chosen to be decreasing and, following previous conventions, we set $v(r) = 1/\log(r + 1)$. With this setting, we now define the expected source side utility for member $s$ and query $q$ as

$$U^s_q = \sum_{d=1}^{D_q} \sum_{r=1}^{m} u^q_{s,d} P^q_s(d, r) v_r = u^T_s P^q_s v.$$

We note that the rows and columns of the ranking policy matrix $P^q_s$ sum to at most one and thus the utility-maximizing probabilistic ranking policy can be found by solving the optimization problem

$$\begin{aligned} \text{arg max}_{P^q_s} \; & u^T_s P^q_s v \\ \text{s.t.} & \sum_{i=1}^{m} P^q_s(d, i) \leq 1 \text{ for all } d \in [D_q], \\ & \sum_{i=1}^{D_q} P^q_s(i, r) = 1 \text{ for all } r \in [m], \\ & 0 \leq P^q_s(i, r) \leq 1 \text{ for all } i \in [D_q], j \in [m]. \end{aligned} \tag{5.1}$$

Here, the first set on constraints ensures that each destination member is suggested in at most one slot and the second set of constraints requires that each slot of the recommendation is filled with exactly one destination member. We note that the optimization problem is linear in $D_q \times m$ variables and can thus be solved with standard methods.

If $D_q = m$, the inequalities in the first set of constraints become equalities and the ranking matrix $P_s^q$ is doubly stochastic. (Singh and Joachims, 2018) use a Birkoff-von-Neumann decomposition to retrieve the deterministic ranking in this case. However, industry applications generally observe settings with $D_q >> m$, i.e. only a very small subset of all possible members is actually ranked for connection recommendation. We generate rankings one recommendation slot at a time by iterating over the columns of the ranking policy matrix $P_s^q$ and selecting a row member randomly according to the probabilities denoted in the column. At each iteration step, the rows corresponding to destination members selected for previous recommendation slots are removed and the column values are rescaled to sum to probability one before sampling a new destination member.

### 5.3.2 Adding destination side fairness

The recommendation procedure described in the previous section optimizes for source side utility without considering the impact on the recommended members which is common practice (Agarwal and Chen, 2016). Yet when only a small subset of members can be recommended for any given query, as is the case for many industry-scale connection recommendation systems, exposure in recommendations becomes a scarce resource that can determine who is able to reconnect with old friends or even who receives job opportunities downstream. In order to understand fairness in recommendations, several metrics have been proposed.

**Demographic parity of exposure**. Among the most commonly proposed metrics is demographic parity of exposure (e.g. Zehlike and Castillo, 2020; Singh and Joachims, 2018; Abdollahpouri et al., 2019a; Singh and Joachims, 2019) which measures the difference in recommendation exposure between groups adjusted for position bias. For a set of disjoint groups of members $G_1, \ldots, G_l$, demographic parity of exposure requires that

$$\frac{1}{|G_k|} \sum_{d \in G_k} \sum_{r=1}^{m} P_s^q(d,r) v_r = \frac{1}{|G'_k|} \sum_{d \in G'_k} \sum_{r=1}^{m} P_s^q(d,r) v_r, \tag{5.2}$$

for all $k, k' \in [l]$ which means that groups are displayed in the recommendations at equal rates. For two groups $G_0$ and $G_1$, this can be compactly written in vector form as $f^T P_s^q v = 0$, where the $d$th entry of $f$ is $f_d = \frac{1(d \in G_0)}{|G_0|} - \frac{1(d \in G_1)}{|G_1|}$. We note that the constraint is linear and can thus be added to the optimization problem in Equation (5.1) without changing the solution approach.

**Dynamic parity of utility**. In addition to demographic parity, we consider a dynamic fairness constraint previously referred to as dynamic parity of utility (Basu et al., 2020). The constraint requires that different groups receive the same rates of expected utility in each recommendation, i.e.

$$\frac{1}{|G_k|} \sum_{d \in G_k} u_{s,d}^q \sum_{r=1}^{m} P_s^d(d,r) v_r = \frac{1}{|G_k'|} \sum_{d \in G_k'} u_{s,d}^q \sum_{r=1}^{m} P_s^d(d,r) v_r, \tag{5.3}$$

for each $k, k'$ and query $q$. For two groups $G_0$ and $G_1$, we can rewrite this as $\tilde{u}_s P_s^q v = 0$ where the $d$th entry of $\tilde{u}_s$ is $(\tilde{u}_s)_d = u_{s,d}^q \left( \frac{1(d \in G_0)}{|G_0|} - \frac{1(d \in G_1)}{|G_1|} \right)$. Note that this is still a linear constraint and can conveniently be added to the optimization problem. In most applications, the relevance $u_{s,d}^q$ is estimated from data that in some way depends on the current state of the recommender system which dynamically changes the constraint over time, e.g. in connection recommender systems we might use the number of existing connections between members. Depending on our understanding of $u_{s,d}^q$, the dynamic parity constraint allows for different interpretations. In our simulation, we will assume that $u_{s,d}^q$ denotes the probability of connection if recommended. In this case, the constraint enforces that destination members in all groups have the same average probability of forming a connection to the source member. If members separate into two groups with shares 2/3 and 1/3 respectively, we would thus expect the recommendation to lead to about twice as many connections to the first group than to the second assuming distributions around the average probability are similar.

### 5.3.3 Scoring model

Connection recommendation requires a notion of relevance of suggestions in order to derive a ranking of members. In our setting, this is captured by the ranking score $u_{s,d}^q$ for source member $s$, destination member $d$ and query $q$ which reflects the utility member $s$ receives from being recommended member $d$ in query $q$. In practice, this utility is often modeled by inserting the probability of connection if recommended, some measure of downstream engagement between the two members or a mixture of the two.

Models for these utility proxies are learned from historic member data and subsequently used to compute ranking scores for new pairs of members. Since the relevant member data is generally not available to researchers, work in this space often relies on deterministic functions for scoring (Basu et al., 2020; Singh and Joachims, 2018). In this chapter, we assume that the likelihood with which a member pair $(s, d)$ connects following a recommendation depends on three main characteristics. First, the larger the current network of $s$, the more likely the member sends invites to recommended members and thus forms connections. This assumption is intuitive since members with large networks tend to be more active in forming connections and in using the platform in general. Second, the more common connections members $s$ and $d$ have, the more likely they are to connect. This is known as triadic closure in the social networks literature and has been shown to be an important predictor in connection forming (Kossinets and Watts, 2006; Liben-Nowell and Kleinberg, 2007; Krackhardt and Handcock, 2007). And third, members with a lot of similarities such as similar demographics, interest, education, workplaces, etc. are generally more likely to connect. This follows the observation that individuals like to be connected to others who are similar to them which is a tendency generally referred to as homophily (McPherson et al., 2001; Louch, 2000; Kossinets and Watts, 2006).

Based on the described components, we assume a model for the connection probability of the pair $(s, d)$ after $d$ has been recommended to $s$ of the form

$$\begin{aligned}\log \frac{p}{1-p} = \beta_0 + \beta_1 \,\text{networkSize}(s) + \beta_2 \,\text{commonConn}(s, d) \\ + \beta_3 \,\text{similarity}(s, d) + \varepsilon,\end{aligned} \tag{5.4}$$

where $\varepsilon \sim \mathcal{N}(0, 0.1)$ is a random noise term. We note that the network sizes and numbers of common connections can be easily computed from the adjacency matrix $A_t$ at time $t$, i.e. $\text{networkSize}(s) = 1^T A_t(s, \cdot)$ and $\text{commonConn}(s, d) = A_t^2(s, d)$. For the similarity between members, we assign each member $i$ in our simulation a fixed covariate vector $X_i$ and then set $\text{similarity}(s, d) = -||X_s - X_d||_2$. All features are scaled to lie in $[0, 1]$. To emulate the noise present in data-driven scoring models, we query the connection probability model once to obtain a ranking score for each member pairing and then again to determine if members chosen to be recommended to each other connect which alters the random noise term affecting the score.

### 5.3.4 Simulation procedure

We simulate connection recommendation in a fixed size graph of $N = 1000$ members with connections evolving over $T = 2500$ discrete time steps. Each recommendation consists of a list of $m = 20$ ranked individuals which are selected trough the probabilistic ranking framework detailed above. The frequency with which recommendations are provided to each member are modeled through an exponential waiting times model dependent on the current network size with mean $\lambda = 1/(0.001 + 0.02 \times$ current network size/1000$)$. Separate experiments are conducted for (1) no fairness intervention, (2) demographic parity of exposure intervention, and (3) dynamic parity of utility intervention while resetting random seeds before each intervention type to ensure equal starting conditions.

Members are separated into 65% majority group (e.g. male members) and 35% minority group (e.g. female members). We independently sample covariate vectors $X_i \in \mathbb{R}^{30}$ for each member $i$ with $X_i \sim \mathcal{N}(\mu_{G_i}, \text{diag}(0.5))$ where $G_i$ denotes the group assigned to member $i$ and $\mu_{G_0}, \mu_{G_1}$ are group-dependent means selected randomly from $U([0,1]^{30})$ and fixed throughout all experiments. The edges of the connection graph are initialized with a stochastic block model with group combination probabilities $(p_0, p_1, p_2)$, i.e. $p_0$ is the probability with which two nodes of group $G_0$ form an initial edge, $p_1$ is the probability with which two nodes of group $G_1$ form an initial edge, and $p_2$ is the probability with which each cross-group pairing forms an initial edge. Group assignments, covariates and the initial graph are fixed for all intervention types in a given simulation run.

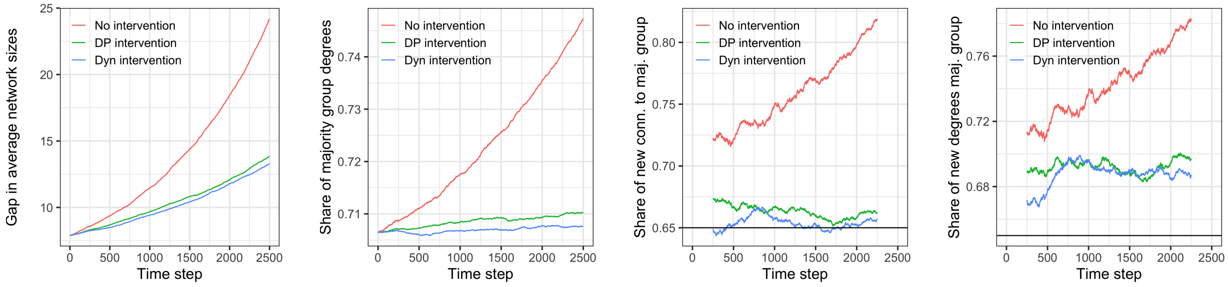
For each intervention type, we iterate over the following steps for each $t = 1, \ldots, T$.

(1) **Select source members:** We select the members with waiting time zero (source members) and decrease the waiting time of other members by one.

(2) **Score member pairings:** For each source member, we compute the relevance scores to all unconnected members in the graph (destination members) by using the scoring model.

(3) **Solve ranking problem:** A ranking of the destination members is obtained by solving the optimization problem in Equation (5.1) subject to fairness constraints if applicable. This is repeated separately for every selected source member.

(4) **Recommendation and addition of connections:** The first $m = 20$ members of each recommendation list are suggested to the source member and a connection is formed based on the probabilities



(a) Absolute difference in average network sizes between groups; 0 if average network sizes are the same.

(b) Share of all degrees that belong to majority group; 0.65 if average network sizes are the same.

(c) Rolling average (window size 500) of majority group share among destination members of new connections; line at 0.65.

(d) Rolling average (window size 500) of majority group share among new degrees; line at 0.65

Figure 5.1: Simulation results over 2,500 time steps for no intervention, demographic parity of exposure intervention (DP) and dynamic parity of utility intervention (Dyn). Results are reported as averages over 10 simulation runs. We see that the increase in network size disparities between groups over time is slowed down but not fully mitigated by the fairness interventions.

obtained by a new call of the scoring model. For this, the probabilities are adjusted for position bias and thresholded at 0.5.

(5) **Update parameters:** As a last step, new waiting times are sampled for the source members in this iteration step based on their new network sizes and we repeat the procedure by returning to step (1).

## 5.4 Empirical results

Experiments are conducted according to the procedure described in Section 5.3.4 for (1) no fairness intervention, (2) demographic parity of exposure intervention, and (3) dynamic parity of utility intervention. Results are reported aggregated over 10 repetitions of the entire simulation procedure.

### 5.4.1 Graph initialization and scoring model

We set the stochastic block model parameters for the graph initialization to $(p_0, p_1, p_2) = (0.04, 0.032, 0.023)$ in order to emulate a realistic setting. This leads to an initial graph in which majority group members have on average 30.16% more connections than members of the minority group, and members from both groups have more common connections with majority group members than minority group members on average. Average similarity between members of the same group is -3.86 while pairings across groups have an average similarity of -4.35. Although initial feature means differ across groups or group pairings, feature distributions heavily overlap and in-group variations outweigh the differences between groups. Figure D.1 summarizes the distribution of ranking features at $t = 0$.

We set the parameters of the scoring model from Equation 5.4 to $\beta_0 = 0, \beta_1 = 50, \beta_2 = 50$ and $\beta_3 = -5$ which has two implications. First, member pairings in the majority group tend to have higher scores than pairings in the minority group because they have larger networks and more common connections, and (2) pairs of members who belong to the same group tend to have higher ranking scores than members from opposite groups because of the different distributions of the number of common connections and the similarity feature. Overall, this leads to decreased ranking scores for minority group members although group membership is not explicitly considered for the computation of scores (Figure D.1).

### 5.4.2 Rich-gets-richer in groups

Figures 5.1 and D.2 depict the results of the connection recommendation experiment if no fairness intervention is applied. Over time, the initial gap in average network sizes between groups increases as majority members are able to grow their networks faster than members of the minority group ( Figure 5.1a). In addition, the majority group share among new connections is increasing over time ( Figures 5.1c and 5.1d) which leads to a superlinear growth of the network size gap and suggests a positive feedback loop which amplifies the advantage of the majority group over time. Starting out with on average 30.18% larger networks, members in the majority group have on average 59.88% more connections after $t = 2500$ time steps. Figure D.2 shows that the distribution of network sizes at $t = 2500$ follows a power law distribution with a particularly long tail for majority group members and lower mode for the minority group suggesting that most majority members have larger networks than most minority members. In addition, the figure depicts the the relation of initial network sizes and network sizes at $t = 2500$ on a log-log scale. The additional curves display the network size of individuals in a counterfactual scenario in which the growth of networks within a group is uniformly distributed among all member of the group, i.e. the curves correspond to $f(\text{network size at } t = 0) = (\text{average increase in } G_i) + (\text{network size at } t = 0)$ on the log-log scale. The result suggests that members who grow their networks more than the average within their group in the given time frame tend to be the members who had larger networks to begin with.

In summary, our key findings from simulation with no fairness intervention are: (1) Unconstrained connection recommendation increases the initial disparity in average network sizes befitting the already advantaged majority population. (2) Network sizes tend to a power law distribution with a lower mode for the minority population. (3) The members whose network sizes are in the tail of the power law distribution tend to be the members who had large networks as compared to the rest of their groups to begin with. Overall, these observations confirm a 'rich-get-richer' or Matthew effect where majority members appear to benefit more from the phenomenon than minority members.

### 5.4.3 Demographic parity of exposure intervention

We conduct the same experiments with demographic parity fairness intervention and present the results in Figures 5.1 and D.3. While majority group members are overexposed with an exposure share of 75.5% without intervention, the demographic parity intervention leads to a majority group exposure share of 66.3% averaged over all time steps and simulation runs which is close to the 65% population share of the majority group. While this suggests that the intervention fulfills its purpose in aggregate, Figures 5.1a and 5.1b show that network sizes are not converging as intended. Although the intervention leads to less outcome disparity than in the unconstrained setting, both the gap in average network sizes and the share of degrees that belong to the majority group increase over the period of the experiment. This is because (1) majority group members in our experiment seek out connection recommendations more frequently, and (2) majority members have higher ranking scores and higher likelihoods to connect. Both of these points lead to more than 65% of new connections being formed to and from majority group members (Figures 5.1c and 5.1d) which exacerbates the differences in network sizes instead of mitigating them. On a high level, the demographic parity of exposure intervention ensures that members of different groups are displayed in recommendations at the same rates which does not lead to equal connection rates when the underlying relevance distributions vary. At $t = 2500$, network sizes in both groups appear to follow a power law distribution with lower mode in the minority group (Figure D.3) which means that the majority of members is still disadvantaged. As compared to no intervention, the average network size of minority group members at $t = 2500$ increases by 2.11 with median increase of 1.

Overall, the results show that enforcing demographic parity of exposure in recommendation lists is not sufficient in order to achieve parity of average network sizes between groups. Although seemingly fair in aggregate, the gap in average network sizes is still increasing over time. This growth is happening at a much slower rate than without fairness intervention suggesting that (part of) the bias amplification feedback loop in the dynamic system is mitigated. We theoretically examine the workings of the intervention in a stylized model in Section 5.5.

### 5.4.4 Dynamic parity of utility intervention

Results for the dynamic parity of utility case are depicted in Figure 5.1 and D.3. The distribution of network sizes after $t = 2500$ time steps of the experiment closely resembles the distribution for the demographic parity of exposure case with generally lower network sizes in the minority group and a median increase of 1 connection as compared to the results of the unconstrained connection recommendation. On average, minority group members gain 2.6 more connections than without intervention.

We observe that the absolute gap in average network sizes and the share of majority group degrees are increasing. However, the increase is happening at a slower rate than in the demographic parity of exposure setting. Figure 5.1c shows that the majority group share among the destination members of new connections hovers around the desired 0.65 mark (on average 0.654) which suggests that the intervention successfully ensures that members of both groups have about the same average probability to gain connections through being displayed in recommendations to other members. However, the majority group share among all new degrees exceeds the desired share and averages to 0.687 over all simulation runs and time steps (Figure 5.1d) which leads to an increasing gap in average network sizes. This is because (1) majority group members seek out recommendations more frequently based on their larger network sizes, and (2) source members who belong to the majority group are able to connect to more of the recommended members since they generally have larger scores and the parity of utility within a single recommendation list does not imply parity of total utilities between recommendation lists for different source members. While the dynamic parity of utility intervention solves some of the problems we observed with the demographic parity of exposure intervention, it suffers from the same limitations regarding the biases introduced by the source side of the connection recommendation system. Our theoretical derivations in Section 5.5 suggest that the dynamic parity of utility intervention can lead to stably fair average network sizes in setting with no source side bias.

## 5.5 Theoretical characterization

### 5.5.1 Urn models and mixed preferential attachment

**Urn models for dynamic systems of unfairness**. The results of the simulation study in Section 5.4 demonstrate how fairness intervention in connection recommender systems can lead to unanticipated long-term effects. In order to understand why these effects occur and how to reach a fair balance of network sizes, we seek out a theoretical analysis of the impact of intervening on the connection recommendation dynamics. While our simulation setup resembles the workings and data setting of real-world connection recommender systems, it is quite complex and a full theoretical characterization of the behavior of the system requires major simplifications to the model structure. Urn models present a class of models whose behavior is more tractable to analyze in theory but that still proves flexible enough to lend itself to various applications (Pemantle, 2007). They have been previously used in the algorithmic fairness literature to model feedback loops in predictive policing (Ensign et al., 2018a).

For the connection recommendation purpose, we employ a type of dynamic growth urn which is also known as preferential attachment model (Barabasi and Albert, 1999). Preferential attachment models are generative random network models which rely on a local growth rule that renders vertices that already have a large number of connections likely to accumulate more connections over time similar to the setting in our simulation study. In an effort to extend the preferential attachment setting to social networks with members of different groups, researchers have proposed a mixed preferential attachment model that allows for a majority-minority partition and consideration of homophily (Avin et al., 2015; 2020). We draw on this type of model to theoretically characterize the workings of fairness intervention in the connection recommendation setting and formally define the mixed preferential attachment model specification we use in the following. Note that, while the models considered in the empirical and theoretical parts of this work are not the same, they are similar enough to warrant the expectation that some of the qualitative observations from the analysis of the mixed preferential attachment model can be translated to insights into the behavior of the realistic simulation study on a group-aggregate level. Appendix D.2 summarizes the similarities and differences between the two models.

**Mixed preferential attachment (MPA) model**. Let $\mathcal{G}_t(r, d_0, \pi)$ be a bi-populated evolving random graph with nodes in groups $G_0$ and $G_1$. Here, $d_0 \in \mathbb{N}$ is the sum of all degrees at $t = 0$, $r \in (0, 0.5]$ is the arrival rate of the $G_1$ vertices, and $\pi \in \mathbb{R}^{2\times 2}$ is the mixing matrix. For simplicity of proofs, we assume the fraction of initial vertices that belong to group $G_1$ is $r$. We denote degree of a vertex $v$ at time $t$ as $d_t(v)$, the sum of all degrees as $d_t$, and the sum of all degrees within groups $G_0$ and $G_1$ as $d_t(G_0)$ and $d_t(G_1)$ respectively. Note that each connection in the graph translates to two degrees, one for the node at either side of the connection. The generative process of the graph works as follows. In each iteration $t$, a new node, which belongs to group $G_1$ with probability $r$ and to $G_0$ otherwise, is added to the network. The new node can be interpreted as the source member who seeks out a connection recommendation and is subsequently connected to exactly one existing node in a two-stage recursive procedure. First, we sample a tentative neighbor at random with probabilities proportional to the degrees of the existing nodes at time $t$, i.e. $P(\text{select node } v) = d_t(v)/d_t$. This member corresponds to the recommended destination member. Second, we denote the groups of the new and selected nodes and sample whether the connection is successful based on the mixing matrix

$$\pi = \begin{bmatrix} p_0 & 1-p_0 \\ 1-p_1 & p_1 \end{bmatrix}.$$

If both nodes belong to group $G_0$, the connection is successful with probability $p_0$. If the first node belongs to group $G_0$ and the second node is from group $G_1$, the connection is successful with probability $1 - p_0$, etc. This step can be interpreted as the members' reaction to the recommendation where the row vectors $\pi_i = (p_i, 1 - p_i)$ represent the homophily preferences of the groups, i.e. how much members prefer connections within their own group over connections to members of the other group. When $p_i \in (0.5, 1)$, members are assumed to be positively biased towards connections within their own group (homophily), and for $p_i \in (0, 0.5)$, members prefer connections to the other group (heterophily). If the connection fails, a new recommendation is made by sampling a new tentative neighbor and repeating the procedure until the new node connects to exactly one existing node.

### 5.5.2 Rich-gets-richer in groups

It is well known that the degree distribution of nodes in preferential attachment models tends to a power law distribution leading to a 'rich-get-richer' phenomenon (e.g. Chung and Linyuan, 2006). The authors of (Avin et al., 2020) extend this result to social network settings where each node belongs to one of two groups. Let $\alpha_t = d_t(G_1)/d_t$ be the rate of minority degrees in a mixed preferential attachment network at time $t$. Then, the paper shows that there is a limit $\alpha$ independent of the initial graph such that $\lim_{t\to\infty} \mathbb{E}[\alpha_t] = \alpha$. Letting $m_{k,t}(G_i)$ denote the number of vertices in group $i$ with degree $k$ at time $t$ and $M_k(G_i) = \lim_{t\to\infty} \mathbb{E}[m_{k,t}(G_i)]/t$, the paper follows that the numbers of degrees in groups tend to power law distributions $M_k(G_i) \propto k^{-\beta(G_i)}$ with different exponents

$$\beta(G_0) = 1 + \frac{1}{c_0} \quad \text{and} \quad \beta(G_1) = 1 + \frac{1}{c_1},$$

where

$$\begin{aligned} c_0 &= \frac{1}{2}\left(\frac{(1-r)p_0}{p_0+\alpha-2p_0\alpha} + \frac{r(1-p_1)}{1-p_1-\alpha+2p_1\alpha}\right), \text{ and} \\ c_1 &= \frac{1}{2}\left(\frac{(1-r)(1-p_0)}{p_0+\alpha-2p_0\alpha} + \frac{rp_1}{1-p_1-\alpha+2p_1\alpha}\right). \end{aligned}$$

Both exponents are functions of the rate of the minority group $r$, the limit $\alpha$ and the mixing matrix $\pi$. In the case of perfect homophily $p_0 = p_1 = 1$ or no homophily bias $p_0 = p_1 = 0.5$, the exponents $\beta(G_i)$ are the same for both groups and no group is disadvantaged in the long term. In the more realistic setting $p_0, p_1 \in (0.5, 1)$, members are more likely to connect to members in the same group but have a non-zero probability of connecting to the members of the other group if recommended. If $p_0 \geq p_1$ in this setting, one can show that $\beta(G_0) \geq \beta(G_1)$ and thus the networks of majority group members outgrow the networks of minority group members in the long run. If $p_1 > p_0 > 0.5$, the picture is more complex, and networks of the minority group can outgrow the networks by the majority group in cases in which the difference between $p_1$ and $p_0$ is large or the rate of the minority group $r$ is close to 0.5.

In the setting corresponding to our simulation study, we have $r = 0.35$ and $p_0 = p_1 > 0.5$ since the similarity feature in scoring model uses the same parameter for both groups. The MPA model suggests that in this case the network sizes in groups tend to a power law distribution with larger networks in the majority group which aligns with our empirical observations. We set $r = 0.35$ in the MPA model and compute the analytical limits of the expected share of minority group degrees for different combinations of $p_0$ and $p_1$. The results in Figure 5.2 (left plot) confirm that (1) most combinations of $p_0$ and $p_1$ lead to

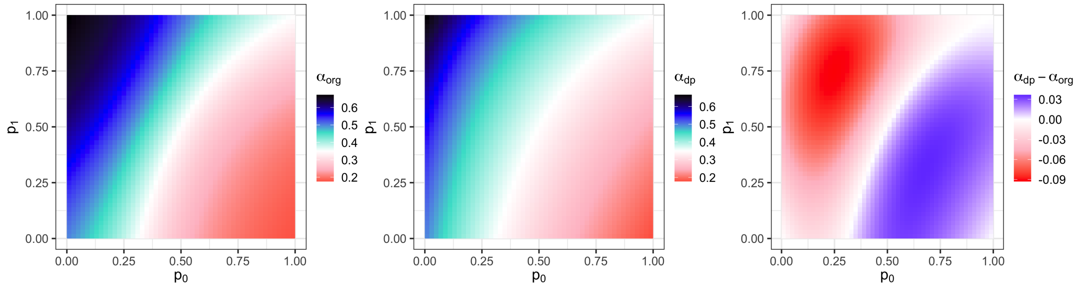


Figure 5.2: Analytical limits $\alpha = \lim_{t\to\infty} \mathbb{E}[\alpha_t]$ for $r = 0.35$ and $p_0, p_1 \in (0, 1)$. In the left and middle plots, $\alpha = r = 0.35$ (white) indicates that the average network size in $G_0$ and $G_1$ is the same in the long term, $\alpha < 0.35$ (pink& red) indicates that members of group $G_0$ have larger average networks and $\alpha > 0.35$ (blue& black) that members of $G_1$ have larger networks. The left plot depicts the limit in the original mixed preferential attachment model, the middle plot the limit with demographic parity intervention, and the right plot the difference between the two. We see that network sizes in groups diverge with and without intervention for most combinations of $p_0$ and $p_1$. While the demographic parity intervention does not correct the entire outcome disparity, it does lead the limit $\alpha$ closer to $r$ as depicted in the right plot. The homophily setting with $p_0, p_1 \in (0.5, 1]$ corresponds to the upper right quadrants of the plots.

divergent network sizes in the long run, and (2) in the homophily setting with $p_0, p_1 > 0.5$ the majority group is more likely to benefit while in the heterophily setting with $p_0, p_1 < 0.5$ the minority group is more likely to be advantaged. For our setting with $p_0 = p_1 > 0.5$, the figure shows that the minority group share of all degrees in the network remains smaller than the desired 35% in the long run which means minority group members are left smaller networks in general.

### 5.5.3 Enforcing parity in recommendations

We saw in our empirical results that enforcing demographic parity of exposure slows down the increase of network size disparities but does not lead to similar distributions of network sizes between groups over time. To understand why this is the case, we analyze the effect of a parity of exposure intervention in the MPA model. The demographic parity of exposure condition in Equation (5.2) requires that the average expected exposure in rankings is the same for members of both groups. Since the MPA model recommendations only have one slot for each query, there is no consideration of position bias in this setting and $v_1 = 1$. The fairness condition becomes

$$\frac{1}{|G_0|} \sum_{d \in G_0} P(\text{recommend } d) = \frac{1}{|G_1|} \sum_{d \in G_1} P(\text{recommend } d).$$

Recall that in the original setting of the MPA model, the incoming source member at time $t$ belongs to minority group $G_1$ with probability $r$, and the recommended destination member is selected from the existing graph with probability determined by their network size $P(\text{select node } v) = d_t(v)/d_t$. In order to fulfill the fairness constraint, we alter the mechanism with which the destination member is sampled and first, sample the group from which to choose a member $G_i$ with $i \sim \text{Bin}(r)$, and second, sample a destination member from that group according to their network sizes with $P(\text{select node } v) = d_t(v)/d_t(G_i)$ for all $v \in G_i$. As before, the two members are connected with probabilities determined by the mixing matrix and the procedure is repeated until the source member forms exactly one connection. We characterize the limiting rate of minority group degrees in this setting in the following theorem. A proof can be found in Appendix D.3.

**Theorem 14.** *Assume a mixed preferential attachment model with demographic parity intervention as described. Then for t $\to \infty$, the share of degrees of the minority group $\alpha_t = d_t(G_1)/d_t$ tends to a fixed value $\alpha = \alpha(\pi, r)$ which is independent of the initial graph. Specifically,*

$$\alpha := \lim_{t\to\infty} \mathbb{E}[\alpha_t] = \frac{1}{2}\left( r\frac{rp_1}{1-r-p_1+2rp_1} - (1-r)\frac{(1-r)p_0}{r+p_0-2rp_0} + 1 \right).$$

If the demographic parity intervention would lead to the same average network sizes in groups in the long-term, we would have $\lim_{t\to\infty} \mathbb{E}[\alpha_t] = r$ which is exactly the fraction of $G_1$ nodes in the graph. However, the limit in Theorem 14 does not equal $r$ in general. A sufficient condition for $\alpha = r$ is the trivial case with $r = 0.5$ and $p_1 = p_0$ in which no bias from the in-group connection probability and the different group sizes is introduced in the first place. However, one can show that in the homophily regime of $p_0, p_1 > 0.5$ and with $r \in (0, 0.5)$, the case $p_0 >= p_1$ leads to $\alpha < r$ and the case $p_1 > p_0$ can lead to $\alpha > r$ if $p_1 - p_0$ is large or $r$ is close to 0.5 similar to the results without intervention. Figure 5.2 displays the limit $\alpha$ after demographic parity intervention for different combinations of $p_0$ and $p_1$ at $r = 0.35$ (middle plot). We see that for only very few combinations of $p_0$ and $p_1$ the limit aligns with the desired share $r = 0.35$. In fact, the left plot shows that those combinations align exactly with the values that lead to fair outcomes even without intervention. Although the demographic parity intervention cannot fully neutralize the bias in network sizes, the figure also shows that it moves the solution $\alpha$ closer to the desired solution $r$ which diminishes the limiting gap in average network sizes. On a high level, the demographic parity intervention can only ensure that members from different groups are recommended at the same rates which is not sufficient to address the asymmetries introduced by the different network sizes and probabilities to connect after recommendation within and in-between groups. The intervention essentially interrupts the group-wise feedback loop stemming from faster growing networks in one of the groups while leaving other sources of bias untouched.

These theoretical results align with the empirical results from the simulation study which show that demographic parity intervention does not lead to equality in average network sizes over time, and minority group members are still largely disadvantaged. Recall that the regime of the simulation study translates to $r = 0.35$ and $p_0 = p_1 > 0.5$ to see this. A key difference between the main simulation and the MPA model are the ground truth probabilities of connection after recommendation. In the MPA model, these probabilities are fixed constants for each of the four combinations of source and destination groups while the simulation model relies on probabilities which are positively affected by larger networks of the source and destination members. In both models, the demographic parity intervention corrects the group-wise feedback loop from larger networks to a higher chance at being selected for recommendation. However, the feedback loop in the main simulation does not only affect the chance of being recommended but also the chance of connection after recommendation which is unaffected by the intervention. This could explain why the disparity growth in Figure 5.1 appears to remain superlinear even with fairness intervention.

### 5.5.4 Stable equality of average network sizes

Both the main simulation study and the analysis of the MPA model show that enforcing demographic parity of exposure in connection recommendation lists is generally not sufficient to reach an equilibrium of equal average network sizes between groups. In the following, we explore what type of intervention is needed to reach this parity state. We recall that in order for the average network sizes to be equal between the two groups in the long-term, the limiting share of minority group degrees $\alpha = \lim_{t\to\infty} \mathbb{E}[\alpha_t]$ needs to be $r$. An easy way to achieve this would be to effectively set the probability of cross-group connections to zero by introducing an additional rejection sampling step that skips a tentative neighbor and resamples whenever the group differs from the group of the source member. Of course in practice, a solution which only allows in-group connections is undesirable. Yet the same rejection sampling idea can be used to derive a more desirable intervention mechanism if we allow for the adjustments to be dynamic.

Consider the following alteration of the mixed preferential attachment iteration step described in Section 5.5.1. Like before, a new member enters the graph and with probability $r$, the new member belongs to group $G_1$. We then select a tentative destination member from the graph with probabilities determined by the network sizes of the existing members. Now, we insert a new rejection sampling step in which the tentative destination member is retained with probability $q_{ij}$ where $G_i$ is the group of the source member and $G_j$ the group of the destination member. If the proposal is rejected, a new tentative destination member is selected as before until a member can be retained successfully. Only then is the destination member recommended to the source, and an edge is inserted with probability determined by the mixing matrix $\pi$. If the connection is unsuccessful, the whole procedure is repeated until the sampled source member connects to exactly one existing member. The following theorem characterizes how the probabilities $q_{ij}$ have to be selected in order to obtain a stable equality of average network sizes between groups. A proof for the theorem is given in Appendix D.4.

**Theorem 15.** *Assume a mixed preferential attachment model with additional rejection sampling step as described above. Let $q_{ij}$ denote the probability with which we retain a tentative destination member of group $G_j$ as possible connection recommendation to a source member of group $G_i$. For each iteration step t, we set*

$$q_{00} = \frac{(1-r)(\alpha_t(p_0-2)+2)}{p_0(\alpha_t - r)}, \qquad q_{01} = 1 - q_{00},$$
$$q_{11} = \frac{(1-\alpha_t)(1-p_1)r}{\alpha_t(p_1-r) - p_1 r + r}, \qquad q_{10} = 1 - q_{11},$$

*and map values outside of $[0,1]$ to $0$ and $1$ respectively. The $q_{ij}$ are selected such that the sampled source member connects to a member of group $G_1$ with probability $r$ and to a member of group $G_0$ with probability $1-r$ in every iteration step t, and thus it holds that $\lim_{t\to\infty}\mathbb{E}[\alpha_t] = r$.*

An important insight from this result is that, in order to ensure stable equal average network sizes between groups in a realistic fashion, we require a dynamic type of fairness intervention that changes with the state of the system. In our case, the probabilities with which a tentative recommendation needs to be rejected depends on the share of minority group degrees in the network $\alpha_t$ and changes as this share evolves. The dynamic intervention is fundamentally different from the demographic parity intervention considered previously. While the demographic parity constraint ensures that members from both groups have the same average probability of being recommended to a source member, the dynamic parity procedure arranges that the probability of connecting to a destination member of the minority group $G_1$ is always $r$ all things considered. This balances out the potential biases introduced on the destination side of the recommendation.

In our experimental setup, an intervention with the described effect is given by the dynamic parity of utility case. Like the rejection sampling idea in the MPA model, dynamic parity of utility ensures that the probability of connecting to members of both groups is proportional to the population group shares in any given recommendation list. Different to the MPA model, additional bias is introduced through the source side in our simulation framework which more closely resembles real-world settings. The theory presented here suggests that dynamic parity of utility intervention can lead to stably fair average network sizes in setting with no source side bias and similar score distributions across groups.

## 5.6 Discussion

**Findings**. We analyze long-term dynamics of fairness intervention in connection recommender systems by (1) studying a simulation-based recommender system patterned after the systems employed by web-scale social networks, and (2) theoretically analyzing how certain interventions on fairness impact the bias amplification dynamics in stylized connection recommendation using mixed preferential attachment models.

Our empirical and theoretical findings suggest that unconstrained connection recommendation leads to amplification of initial differences in average network sizes between groups, and a group-wise rich-get-richer effect benefiting the majority population and especially majority group members who had relatively large networks to begin with which is in line with previous research (Chen et al., 2020; Yao et al., 2021). We find that intervening by enforcing demographic parity of exposure in recommendation lists as commonly suggested in the literature (e.g. Zehlike and Castillo, 2020; Singh and Joachims, 2018; Abdollahpouri et al., 2019a; Singh and Joachims, 2019) leads to less bias amplification but is not sufficient in order to mitigate an increase in the disparities in network sizes over time. Although seemingly 'fair' in aggregate, most minority group members remain disadvantaged in the long run. Moving to dynamic parity of utility intervention alleviates some of the problems posed by the demographic parity of exposure case but still results in increasing disparities over time. This is because fairness is only evoked in individual recommendation lists and does not affect bias introduced through the source side of the recommendations.

Most commonly, the efficacy of fairness intervention in recommendation is measured by a single fixed fairness criterion that is evaluated in a one-shot or time-aggregate manner (Singh and Joachims, 2018; Zehlike et al., 2017; Zehlike and Castillo, 2020). Our work demonstrates how this can lead to deployment of fairness enhancing algorithms with unforeseen consequences in the long run by hiding variations in fairness and other metrics over time. Ultimately, connection recommendation operates on a dynamical system which needs to be taken into account explicitly in order to ensure equitable outcomes in the long run.

**Sensitivity to source side bias**. Theoretical analysis of our urn-based model suggests that dynamic parity of utility intervention can mitigate disparities in network sizes if and only if no bias is introduced through the source side of the recommendation process. Yet, source side bias is to be expected in real-world settings and our simulation study opts to incorporate such bias in several ways: (1) We assume that users with larger networks are served connection recommendations more frequently. (2) We assume that users with larger networks are more likely to form connections based on recommendations. (3) We assume initial differences in the distributions of users' similarity and their number of common connections.

The last point is build around the homophily and triadic closure ideas discussed in Section 5.3.3 and leads to more source utility and thus more new connections per recommendation for the majority group. The first two mechanisms follow a more intuitive rationale. We generally assume a positive correlation between network size and platform activity levels of users which naturally leads users with larger networks to come across more recommendations for connections. Given that those users have large networks, we assume that they are somewhat proactive in forming ties which leads them, on average, to form more connections per recommendation than users with smaller networks. Both of these assumptions have to be carefully checked in practice and might not hold true in all settings, e.g. one could imagine a scenario in which a user with very large network stops to proactively seek out connections based on recommendations instead relying on connection invitations from others. However, as long as some sort of source side bias between groups is introduced, our findings suggest that the studied types of fairness interventions are not sufficient to prevent outcome disparities and targeted source side fairness interventions are needed.

**Moving towards real-world impact**. Most research in fair recommendation abstracts away from specific application settings which has been criticized as ineffective and in some cases even harmful (Patro et al., 2022; Selbst et al.). Instead, our work assumes a concrete connection recommendation setting patterned after the systems employed by real-world social networks which allows us to draw concrete conclusions for application. Nevertheless, assessing the exact impact and possible side effects of fairness intervention after deployment in real-world systems remains difficult because of additional complexity and noise (Holstein et al., 2019). One component that often remains unaddressed is the role of user feedback. In many cases, destination side recommendation utility and fairness are measured by using ranking exposure as a proxy variable which ignores potential variations in user response (Patro et al., 2022). In settings where users directly discriminate against one group, increasing the group's exposure in recommendations as part of a fairness enhancement effort could even lead to adverse outcomes for the group as demonstrated in other settings (Liu et al., 2018; Agarwal et al., 2014). Understanding and modeling the role of human biases in connection recommendation is an integral extension of the work presented here. Further complications are introduced by the noise and uncertainty in real-life recommendation settings. For example, relevance scores can usually only be estimated from data plagued by selection bias which introduces noise and additional biases into the system (Emelianov et al., 2020; Patro et al., 2022). In order to obtain balanced data, we would have to employ a uniformly at random recommendation policy which has been attempted by researchers in the past (Jiang et al., 2019; Liu et al., 2020) but inevitably hurts the experience of members (Chen et al., 2020). In addition, most fairness measures assume access to individual level demographic information which can be hard to obtain in practice for legal reasons and concerns around privacy (Holstein et al., 2019; Bogen et al., 2020; Andrus et al.; Patro et al., 2022).

**Scalability in industry applications**. Web-scale ranking algorithms are required to balance recommendation performance and personalization with scalability in the number of visits, the number of items to be ranked, the amount of training data, etc. (Agarwal et al., 2014). To this end, algorithms often target several engagement metrics at once. In the connection recommendation setting, the target utility proxy could be a mixture of models of the probability that a connection invite is send, the probability that an invite would be accepted, and measures of down-stream engagement. Multi-objective optimization provides an efficient way to derive recommendation policies which balance different business interests in real-world recommender systems (Agarwal et al., 2011; 2018b). In our work, we compute relevance scores for connection recommendation based on a key engagement metric and intervene on fairness by re-ranking the obtained ordering. Large-scale recommendation systems usually favor post-processing strategies like this over pre- or in-processing fairness intervention as they are typically agnostic to the underlying model structures and scale well with large amounts of data and across similar applications (Geyik et al., 2019; Nandy et al., 2020). We note that our simulation study invokes fairness criteria by directly solving separate optimization problems for each session which can lead to latency in online settings. This problem can be solved by solving an aggregate primal optimization problem and subsequently relying on a dual trick to quickly obtain re-rankings online (Basu et al., 2020).

**Homophilic behavior and algorithmic glass ceilings**. We find that connection recommender systems can exacerbate disparities between different demographic groups of users leading to a group wise rich-get-richer effect that benefits the majority population. This finding aligns with the observations of previous research in the social networks area. (Stoica et al., 2018) study the impact of gender and homophily in the context of Instagram recommendations. They authors demonstrate the existence of an algorithmic glass ceiling that prevents equal representation of women and people of color based on reinforced pre-existing disparities. (Hofstra et al., 2017) analyze patterns of segregation based on gender and ethnicity on Facebook, and (Karimi et al., 2018) focus on the effects of homophilic behavior paired with a group size difference in preferential attachment models. The study finds that smaller minority groups suffer more from homophily. Related phenomenons have been described as the filter bubble problem of link prediction (Masrour et al., 2020; Nguyen et al., 2014) which is a term used more generally to describe how algorithmic personalization can lead to overly homogeneous recommendations essentially isolating users from different viewpoints. Lastly, research in social psychology finds that homophilic tendencies hinder women in male dominated fields from forming professional connections which can lead to sex segregation over time (Roth, 2004), and is closely related to the idea of tokenism (Davis, 2016; Kanter, 1977).

# Chapter 6

# Bibliography

Himan Abdollahpouri and Masoud Mansoury. Multi-sided exposure bias in recommendation. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2006.15772.

Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Beyond personalization: Research directions in multistakeholder recommendation. *arXiv preprint*, 2019a. URL https://arxiv.org/abs/1905.01986.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. *arXiv preprint*, 2019b. URL https://arxiv.org/abs/1907.13286.

Roy Adams, Yuelong Ji, Xiaobin Wang, and Suchi Saria. Learning models from data with measurement error: Tackling underreporting. In *International Conference on Machine Learning (IMCL '20*, 2019.

Julius A Adebayo et al. *FairML: ToolBox for diagnosing bias in predictive modeling*. PhD thesis, Massachusetts Institute of Technology, 2016.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17 (6):734–749, 2005.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, 2018a.

Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, and Xuanhui Wang. Click shaping to optimize multiple objectives. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 132–140, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308137.

Deepak Agarwal, Bee-Chung Chen, Rupesh Gupta, Joshua Hartman, Qi He, Anand Iyer, Sumanth Kolar, Yiming Ma, Pannagadatta Shivaswamy, Ajit Singh, and Liang Zhang. Activity ranking in linkedin feed. KDD '14, page 1603–1612, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569.

Deepak Agarwal, Kinjal Basu, Souvik Ghosh, Ying Xuan, Yang Yang, and Liang Zhang. Online parameter selection for web-based ranking problems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 23–32, New York, NY, USA, 2018b. Association for Computing Machinery. ISBN 9781450355520.

Deepak K. Agarwal and Bee-Chung Chen. *Statistical Methods for Recommender Systems*. Cambridge University Press, 2016.

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. The challenge of imputation in explainable artificial intelligence models. *arXiv preprint, arXiv:1907.12669*, 2019.

Dennis J. Aigner and Glen G. Cain. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, 30(2):175, 1977.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 2021.

Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. Long-term dynamics of fairness intervention in connection recommender systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 2022a.

Nil-Jana Akpinar, Manish Nagireddy, Logan Stapleton, Hao-Fei Cheng, Haiyi Zhu, Steven Wu, and Hoda Heidari. A sandbox tool to bias(stress)-test fairness algorithms. *arXiv preprint*, 2022b. URL https://arxiv.org/abs/2204.10233.

Nil-Jana Akpinar, Zachary C. Lipton, and Alexandra Chouldechova. The impact of differential feature under-reporting on algorithmic fairness. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2401.08788.

Michelle Alexander. *The new Jim crow (10th anniversary edition) the new Jim crow (10th anniversary edition)*. New Press, New York, NY, 10 edition, July 2020.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 4th Conference on Fairness, Accountability, and Transparency (FAccT 2021)*.

J D Angrist and Jorn-Steffen Pischke. *Mostly harmless econometrics*. Princeton University Press, Princeton, NJ, December 2008.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Nicholas C Arpey, Anne H Gaglioti, and Marcy E Rosenbaum. How socioeconomic status affects patient perceptions of health care: A qualitative study. *J. Prim. Care Community Health*, 8(3):169–175, July 2017.

Edem F Avakame, James J Fyfe, and Candace McCoy. "did you call the police? what did they do?" an empirical assessment of black's theory of mobilization of law. *Justice Quarterly*, 16(4):765–792, 1999.

Chen Avin, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. Homophily and the glass ceiling effect in social networks. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, January 2015.

Chen Avin, Hadassa Daltrophe, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. Mixed preferential attachment model: Homophily and minorities in social networks.

*Physica A: Statistical Mechanics and its Applications*, 555:124723, Oct 2020. ISSN 0378-4371.

Alexander Babuta and Marion Oswald. *Data Analytics and Algorithms in Policing in England and Wales*. Royal United Services Institute, London, UK, 2020.

T.C. Bailey, M.S. Carvalho, T.M. Lapa, W.V. Souza, and M.J. Brewer. Modeling of under-detection of cases in disease surveillance. *Annals of Epidemiology*, 15(5):335–343, May 2005. doi: 10.1016/j.annepidem.2004.09.013. URL https://doi.org/10.1016/j.annepidem.2004.09.013.

Niels Bantilan. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services*, 36(1):15–30, 2018.

Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, October 1999.

Geoffrey C Barnes and Jordan M Hyatt. Classifying adult probationers by forecasting future offending. *National Institute of Justice. Retrieved February*, 4:2020, 2012.

Francisco Barreras, Alvaro Riascos, and Mónica Ribero. Comparison of different crime prediction models in bogotá. 2016.

Alexander Bartik and Scott Nelson. Deleting a signal: Evidence from pre-employment credit checks. *SSRN Electronic Journal*, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3498458. URL https://bfi.uchicago.edu/wp-content/uploads/BFI_WP_2019137.pdf.

Kinjal Basu, Cyrus DiCiccio, Heloise Logan, and Noureddine El Karoui. A framework for fairness in two-sided marketplaces. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2006.12756.

Eric P Baumer. Neighborhood disadvantage and police notification by victims of violence. *Criminology*, 40(3):579–616, 2002.

Eric P Baumer and Janet L Lauritsen. Reporting crime to the police, 1973–2005: A multivariate analysis of long-term trends in the national crime survey (ncs) and national crime victimization survey (ncvs). *Criminology*, 48(1):131–185, 2010.

Katherine Beckett, Kris Nyrop, and Lori Pfingst. Race, drugs, and policing: Understanding disparities in drug delivery arrests. *Criminology*, 44(1):105–137, 2006.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Mark T. Berg, Lee Ann Slocum, and Rolf Loeber. Illegal behavior, neighborhood context, and police reporting by victims of violence. *Journal of Research in Crime and Delinquency*, 50(1):75–103, 2011.

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*, 50(1):3–44, July 2018.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2212–2220. Association for Computing Machinery, Jul 2019. ISBN 9781450362016.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. Algorithmic fairness and vertical equity: Income fairness with irs tax audit models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *Proceedings of the 2020 Symposium on the Foundations of Responsible Computing (FORC)*, 2020.

Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 3rd Conference on Fairness, Accountability and Transparency (FAT* 2020)*, 2020.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, 2016.

Stacey J Bosick, Callie Marie Rennison, Angela R Gover, and Mary Dodge. Reporting violence to the police: Predictors through the life course. *Journal of Criminal Justice*, 40(6):441–451, 2012.

P Jeffrey Brantingham, Matthew Valasik, and George O Mohler. Does predictive policing lead to biased arrests? results from a randomized controlled trial. *Statistics and public policy*, 5(1):1–6, 2018.

Hermann Brenner and Dana Loomis. Varied forms of bias due to nondifferential error in measuring exposure. *Epidemiology*, 5(5):510–517, 1994.

Irwin Bross. Misclassification in 2 X 2 tables. *Biometrics*, 10(4):478, December 1954.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**, 2018.

Bradley Butcher, Chris Robinson, Miri Zilka, Riccardo Fogliato, Carolyn Ashurst, and Adrian Weller. Racial disparities in the enforcement of marijuana violations in the us. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 130–143, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471.

Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3995–4004, 2017.

Lauren Bennett Cattaneo and Heidi L. M. DeLoveh. The role of socioeconomic status in helpseeking from hotlines, shelters, and police among a national sample of women experiencing intimate partner violence. *American Journal of Orthopsychiatry*, 78(4):413–422, 2008.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

Danton S. Char, Nigam H. Shah, and David Magnus. Implementing machine learning in health care — addressing ethical challenges. *New England Journal of Medicine*, 378(11):981–983, March 2018.

Robert Cheetham. Why we sold hunchlab. https://www.azavea.com/blog/2019/01/23/why-we-sold-hunchlab/, Jan 2019. Azavea. [Online; accessed 1/20/2021].

Irene Y. Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory? In *NIPS*, 2018.

Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi.

Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4(1):123–144, July 2021.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *arXiv preprint*, 2020. URL http://arxiv.org/abs/2010.03240.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, January 2019.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, June 2017.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, 2020.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *FAT**, 2018a.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency (FAT*)*, 2018b.

Haitao Chu, Zhaojie Wang, Stephen R. Cole, and Sander Greenland. Sensitivity analysis of misclassification: A graphical and a bayesian approach. *Annals of Epidemiology*, 16(11):834–841, 2006.

Fan Chung and Lu Linyuan. *Complex graphs and networks*. American Mathematical Society, Providence, RI, 2006. ISBN 978-0-8218-3657-6.

Federico Cismondi, André S. Fialho, Susana M. Vieira, Shane R. Reti, João M.C. Sousa, and Stan N. Finkelstein. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine*, 58(1):63–72, May 2013.

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the international conference on Web search and web data* *mining - WSDM '08*. ACM Press, 2008.

Maria Cuellar and Maria De-Arteaga. Algoritmos y crímenes, Aug 2020. URL https://www.semana.com/opinion/articulo/prevencion-de-delitos--columnistas-maria-de-arteaga-gonzalez-y-maria-cuellar-co 608182/.

Jessica Dai, Sina Fazelpour, and Zachary Lipton. Fair machine learning under partial compliance. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 55–65, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 2020.

Emmalon Davis. Typecasts, tokens, and spokespersons: A case for credibility excess as testimonial injustice. *Hypatia*, 31(3):485–501, 2016. ISSN 0887-5367.

Cámera de Comercio de Bogotá. Encuesta de percepción y victimización. https://www.ccb.org.co/Transformar-Bogota/Seguridad-y-Justicia/Encuesta-de-Percepcion-y-Victimizacion. [Online; accessed 10/4/20].

Cámera de Comercio de Bogotá. *Encuesta de Percepción y Victimización - Primer semestre de 2014 (Chapinero)*. 2014. [Presentation].

Cámera de Comercio de Bogotá. *Observatorio de Seguridad en Bogotá: Balance de Seguridad en Bogotá - 2014*, volume 48. 2015.

Guilherme Lopes de Oliveira, Rosangela Helena Loschi, and Renato Martins Assunção. A random-censoring poisson model for underreported data. *Statistics in Medicine*, 36(30):4873–4892, October 2017. doi: 10.1002/sim.7456. URL https://doi.org/10.1002/sim.7456.

Matthew Desmond, Andrew V. Papachristos, and David S. Kirk. Police violence and citizen crime reporting in the black community. *American Sociological Review*, 81(5):857–876, 2016.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Mustafa Dosemeci, Sholom Wacholder, and Jay H. Lubin. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology*, 132(4): 746–748, 1990.

Wei Du and Xintao Wu. Robust fairness-aware learning under sample selection bias. *arXiv preprint arXiv:2105.11570*, 2021.

Mateo Dulce, Simón Ramírez-Amaya, and Álvaro Riascos. Efficient allocation of law enforcement resources using predictive police patrolling. *arXiv preprint arXiv:1811.12880*, 2018.

Mateo Dulce Rubio et al. Predicting criminal behavior with levy flights using real data from bogota. Master's thesis, Uniandes, 2018.

Michaela Dvorzak and Helga Wagner. Sparse bayesian modelling of underreported count data. *Statistical Modelling*, 16(1):24–46, June 2015. doi: 10.1177/1471082x15588398. URL https://doi.org/10.1177/1471082x15588398.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*. ACM Press, 2012.

Jessie K Edwards, Stephen R Cole, Melissa A Troester, and David B Richardson. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am. J. Epidemiol.*, 177(9):904–912, May 2013.

Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* 2018)*, 2018.

L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, 2018.

Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit variance. In *Proceedings of the 21st ACM Conference on Economics and Computation*

*(EC 2020)*, 2020.

Zeynep Engin and Philip Treleaven. Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *The Computer Journal*, 62(3):448–460, 2019.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *FAT**, 2018a.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2018b.

Virginia Eubanks. A child abuse prediction model fails poor families, 2018. URL https://www.wired.com/story/excerpt-from-automating-inequality/.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7):3217–3258, March 2021.

Andres Ferraro, Xavier Serra, and Christine Bauer. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, page 249–254. Association for Computing Machinery, Mar 2021. ISBN 9781450380553.

Casey Fiesler, Natalie Garrett, and Nathan Beard. What do we teach when we teach tech ethics? a syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 289–295, 2020.

Dylan Fitzpatrick, Wilpen Gorr, and Daniel B. Neill. Hot-spot-based predictive policing in pittsburgh: A controlled field experiment. *Preprint*, 2018. URL http://halley.exp.sis.pitt.edu/comet/presentColloquium.do?col_id=16153. [Online; accessed 1/20/21].

Dylan J. Fitzpatrick, Wilpen L. Gorr, and Daniel B. Neill. Keeping score: Predictive analytics in policing. *Annual Review of Criminology*, 2(1):473–491, 2019.

Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 100–111, New York, NY, USA, 2021. Association for Computing Machinery.

Centre for Social Data Analytics at the Auckland University of Technology. Implementing the hello baby prevention program in allegheny county. 2020. URL https://www.alleghenycountyanalytics.us/wp-content/uploads/2020/12/Hello-Baby-Methodology-v6.pdf.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Christian Fricke. Missing fairness: The discriminatory effect of missing values indatasets on fairness in machine learning. *Master thesis*, 2020.

Wayne A Fuller. *Measurement Error Models*. Series: Wiley Series in Probability & Mathematical Statistics. John Wiley & Sons, Nashville, TN, August 1987.

Nikhil Garg, Hannah Li, and Faidra Monachou. Dropping standardized testing for admissions trades off information and access, 2020.

Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 445–453. Association for Computing Machinery, Mar 2021. ISBN 9781450382977.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, Philadelphia, PA, 3 edition, November 2013.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery*, KDD '19, page 2221–2231, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11): 1544, November 2018.

Alejandro Giménez-Santana, Joel M. Caplan, and Grant Drawve. Risk terrain modeling and socio-economic stratification: Identifying risky places for violent crime victimization in bogotá, colombia. *European Journal on Criminal Policy and Research*, 24(4):417–431, 2018.

Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system. *ACM Transactions on Management Information Systems*, 6(4):1–19, 2016.

Lovedeep Gondara and Ke Wang. MIDA: Multiple imputation using denoising autoencoders. In *Advances in Knowledge Discovery and Data Mining (PAKDD '18)*. 2018.

Wilpen L. Gorr and YongJei Lee. Early warning system for temporary crime hot spots. *Journal of Quantitative Criminology*, 31(1):25–47, 2014.

Heike Goudriaan, James P. Lynch, and Paul Nieuwbeerta. Reporting to the police in western nations: A theoretical analysis of the effects of social context. *Justice Quarterly*, 21(4):933–969, 2004.

Heike Goudriaan, Karin Wittebrood, and Paul Nieuwbeerta. Neighbourhood characteristics and reporting crime. *The British Journal of Criminology*, 46(4):719–742, 2006.

Martin S. Greenberg and R. Barry Ruback. *After the Crime*. Springer US, 1992.

Sander Greenland. Sensitivity analysis and bias analysis. In *Handbook of Epidemiology*, pages 685–706. Springer New York, 2014.

Brandi A. Greer, James D. Stamey, and Dean M. Young. Bayesian interval estimation for the difference of two independent poisson rates using data subject to under-reporting. *Statistica Neerlandica*, 65(3):259–274, April 2011. doi: 10.1111/j.1467-9574.2011.00483.x. URL https://doi.org/10.1111/j.1467-9574.2011.00483.x.

Rolf H H Groenwold and Olaf M Dekkers. Missing data: the impact of what is not there. *European Journal of Endocrinology*, 183(4):E7–E9, October 2020.

Carmen M. Gutierrez and David S. Kirk. Silence speaks: The relationship between immigration and the underreporting of crime. *Crime & Delinquency*, 63(8):926–950, 2015.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Patricia Y Hashima and David Finkelhor. Violent victimization of youth versus adults in the national crime victimization survey. *Journal of interpersonal Violence*, 14(8):799–820, 1999.

Jerry Hausman. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives*, 15(4):57–67, 2001.

Bas Hofstra, Rense Corten, Frank van Tubergen, and Nicole B. Ellison. Sources of segregation in social networks: A novel approach using facebook. *American Sociological Review*, 82(3):625–656, 2017.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–16, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702.

Kimberly A Houser and Debra Sanders. The use of big data analytics by the irs: Efficient solutions or the end of privacy as we know it. *Vand. J. Ent. & Tech. L.*, 19:817, 2016.

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I. Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539–542, 2018.

Keith L. Hullenaar and R. Barry Ruback. Gender interaction effects on reporting assaults to the police. *Journal of Interpersonal Violence*, 2020.

Priscillia Hunt, Jessica Saunders, and John S. Hollywood. *Evaluation of the Shreveport Predictive Policing Experiment*. RAND Corporation, Santa Monica, CA, 2014.

Rolf Jagerman, Ilya Markov, and Maarten de Rijke. When people change their mind. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, January 2019.

Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User - Adapted Interaction; Dordrecht*, 25(5):427–491, Dec 2015. ISSN 0924-1868.

Fieke Jansen. *Data Driven Policing in the Context of Europe*. DATAJUSTICE project, 2018.

Vincent Jeanselme, Maria De-Arteaga, Zhe Zhang, Jessica Barrett, and Brian Tom. Imputation strategies under clinical presence: Impact on algorithmic fairness. In *Machine Learning for Health*, 2022.

Brian Jordan Jefferson. Predictable policing: Predictive crime mapping and geographies of policing and race. *Annals of the American Association of Geographers*, 108(1):1–16, 2017.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, January 2019.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems*, 25(2):7, 2007.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. *ACM SIGIR Forum*, 51(1):4–11, August 2017.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

Rosabeth Kanter. *Men and Women of the Corporation*. Basic Books, New York, 1977.

Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific Reports*, 8(1), July 2018.

Fereshte Khani and Percy Liang. Feature noise induces loss discrepancy across groups. In *International Conference on Machine Learning (ICML '20)*, 2020.

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 196–205, 2021.

G. King, J. Honaker, A. Joseph, and K. Scheve. Analyzing incomplete political science data: an alternative algorithm for multipleimputation. *Am. Polit. Sci. Rev.95*, pages 49–69, 2001.

Chamari I Kithulgoda, Rhema Vaithianathan, and Dennis P Culhane. Predictive risk modeling to identify homeless clients at risk for prioritizing services using routinely collected data. *Journal of Technology in Human Services*, 40(2):134–156, 2022.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXic preprint arXiv: 1609.05807*, 2017.

Ronny Kohavi and Barry Becker. UCI machine learning repository, adult income dataset, 2017. URL https://archive.ics.uci.edu/ml/datasets/census+income.

Gueorgi Kossinets and Duncan J. Watts. Empirical analysis of an evolving social network. *Science*, 311 (5757):88–90, Jan 2006. ISSN 0036-8075.

David Krackhardt and Mark S. Handcock. Heider vs simmel: Emergent features in dynamic structures. In *Statistical Network Analysis: Models, Issues, and New Directions*, pages 14–27. Springer Berlin Heidelberg, 2007.

Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, and Michael I. Jordan. Do offline metrics predict online performance in recommender systems? *arXiv preprint*, 2020. URL https://arxiv.org/abs/2011.07931.

Marine LeMorvan, Julie Josse, Erwan Scornet, and Gael Varoquax. What's a good imputation to learn missing values? In *Advances in Neural Information Processing Systems (Neurips '21)*, 2021.

Karen Levy, Kyla E Chasalow, and Sarah Riley. Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science*, 17:309–334, 2021.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin M. Marlin. Misgan: Learning from incomplete data with generative adversarial networks. *arXiv preprint, arXiv:1902.09599*, 2019.

Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, 9(1), 2018.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

LinkedIn. People you may know. URL https://engineering.linkedin.com/teams/data/artificial-intelligence/people-you-may-know. [Online; accessed 8/15/22].

Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2020.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning (ICLM 2018)*, 2018.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2007.

Hugh Louch. Personal network integration: transitivity and homophily in strong-tie relations. *Social Networks*, 22(1):45–64, May 2000.

Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

Cynthia A Mamalian and Nancy Gladys La Vigne. *The use of computerized crime mapping by law enforcement: Survey results*. US Department of Justice, Office of Justice Programs, National Institute of Justice, 1999.

C Dianne Martin. The case for integrating ethical and social impact into the computer science curriculum. In *The supplemental proceedings of the conference on Integrating technology into computer science education: working group reports and supplemental proceedings*, pages 114–120, 1997.

Farzan Masrour, Tyler Wilson, Heng Yan, Pang-Ning Tan, and Abdol Esfahanian. Bursting the filter bubble: Fairness-aware network link prediction. *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 34(01):841–848, Apr 2020. ISSN 2159-5399.

Sandra G Mayson. Bias in, bias out. *The Yale Law Journal*, 128(8):2218–2300, 2019.

John F McCarthy, Robert M Bossarte, Ira R Katz, Caitlin Thompson, Janet Kemp, Claire M Hannemann, Christopher Nielson, and Michael Schoenbaum. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the us department of veterans affairs. *American* *journal of public health*, 105(9):1935–1942, 2015.

P.E. McKnight, K.M. McKnight, S. Sidani, and A. J. Figueredo. *Missing Data: A Gentle Introduction*. 2007.

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, August 2001.

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.

Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2243–2251. Association for Computing Machinery, Oct 2018. ISBN 9781450360142.

Amalia R Miller and Carmit Segal. Do female officers improve law enforcement quality? effects on crime reporting and domestic violence. *The Review of Economic Studies*, 86(5):2220–2247, 2018.

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

G. O. Mohler, M. B. Short, Sean Malinowski, Mark Johnson, G. E. Tita, Andrea L. Bertozzi, and P. J. Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411, 2015.

George Mohler. Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497, 2014.

Elías Moreno and Javier Girón. Estimating with incomplete count data a bayesian approach. *Journal of Statistical Planning and Inference*, 66(1):147–159, 1998. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(97)00073-6. URL https://www.sciencedirect.com/science/article/pii/S0378375897000736.

Rachel E Morgan and Barbara A Oudekerk. *Criminal victimization, 2018*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2019.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 429–438. Association for Computing Machinery, Jul 2020. ISBN 9781450380164.

Preetam Nandy, Cyrus Diciccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. Achieving fairness via post-processing in web-scale recommender systems. *arXiv preprint*, 2020. URL http://arxiv.org/abs/2006.11350.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. Exploring the filter bubble. In *Proceedings of the 23rd international conference on World wide web (WWW 2014)*, 2014.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. Fair ranking: A critical review, challenges, and future directions. In *FAccT*, 2022.

Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:1–79, 2007.

Edmund Phelps. The statistical theory of racism and sexism. *American Economic Review*, 62(4):659–61, 1972.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, and Sharad Goel. A large-scale analysis of racial disparities in police stops across the united states. *Nature Human Behaviour*, 4 (7):736–745, 2020.

PredPol. URL https://www.predpol.com/law-enforcement/#predPolicing. [Online;

accessed 10/7/20].

PredPol. Machine learning and policing. https://blog.predpol.com/machine-learning-and-policing, 2017a. [Online; accessed 1/20/21].

PredPol. Proven results of our predictive policing software. https://www.predpol.com/results/, 2017b. [Online; accessed 1/20/21].

Dewi Rahardja and Dean M. Young. Confidence intervals for the risk ratio using double sampling with misclassified binomial data. *Journal of Data Science*, 9(4):529–548, 2021.

Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866, December 2018.

Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.*, 33(3):299–318, 08 2018.

UCI Machine Learning Repository. German credit data. 1994. URL https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

Rashida Richardson, Jason Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems and justice. *New York University Law Review Online*, 192, 2019.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9 (1), 2018.

Louise Marie Roth. The social psychology of tokenism: Status and homophily processes on wall street. *Sociological perspectives: SP: official publication of the Pacific Sociological Association*, 47 (2):189–214, Jun 2004. ISSN 0731-1214.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)*, 19(4):1–26, 2019.

Konstantinos Sechidis, Matthew Sperrin, Emily S. Petherick, Mikel Luján, and Gavin Brown. Dealing with under-reported variables: An information theoretic solution. *International Journal of Approximate Reasoning*, 85:159–177, 2017.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2nd Conference on Fairness, Accountability and Transparency (FAT* 2019)*.

Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. Exploring artist gender bias in music recommendation. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2009.01715.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 workshop: Machine Learning for the Developing World*, 2017.

Aaron Shapiro. Reform predictive policing. *Nature*, 541(7638):458–460, 2017.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, page 2219–2228, New York, NY, USA, 2018. Association for Computing Machinery.

Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems (Neurips 2019)*, volume 32, 2019.

Lee Ann Slocum. The effect of prior police contact on victimization reporting: Results from the police–public contact and national crime victimization surveys. *Journal of Quantitative Criminology*, 34 (2):535–589, 2017.

Lee Ann Slocum, Terrance J. Taylor, Bradley T. Brick, and Finn-Aage Esbensen. Neighborhood structural characteristic, individual-level attitudes, and youths' crime reporting intentions. *Criminology*, 48(4): 1063–1100, 2010.

Daniel Sprick. Predictive policing in china: An authoritarian dream of public security. *Nordic Journal of*

*Law and Social Research*, 2020.

Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. Algorithmic glass ceiling in social networks. In *Proceedings of the 2018 World Wide Web Conference (WWW 2018)*. ACM, 2018.

Oliver Stoner, Theo Economou, and Gabriela Drummond Marques da Silva. A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association*, 114 (528):1481–1492, apr 2019. doi: 10.1080/01621459.2019.1573732. URL https://doi.org/10.1080%2F01621459.2019.1573732.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9. 2021.

Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(10), 2013.

Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation. *Center for Social data Analytics*, 2017.

Marvin Van Bekkum and Frederik Zuiderveen Borgesius. Digital welfare fraud detection and the dutch syri judgment. *European Journal of Social Security*, 23(4):323–340, 2021.

S. Van Buuren and K. Oudshoorn. Flexible multivariate imputation by mice. *Leiden: TNO*, 1999.

David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729.e27, 2018.

Alejandro Veen and Frederic P Schoenberg. Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, 103(482):614–

624, June 2008.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 526–536, 2021.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I Jordan. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343*, 2020.

Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, February 2018.

Yanchen Wang and Lisa Singh. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2):101–119, May 2021.

Adam M Watkins. Examining the disparity between juvenile and adult victims in notifying the police: A study of mediating variables. *Journal of research in Crime and Delinquency*, 42(3):333–353, 2005.

David Weisburd, Rosann Greenspan, Stephen Mastrofski, James J Willis, Police Foundation, and United States of America. Compstat and organizational change: A national assessment. *National Institute of Justice*, 2008.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

Rainer Winkelmann. *Econometric analysis of count data*. Springer, Berlin, Germany, 5 edition, April 2008.

Ali Winston. Palantir has secretly been using new orleans to test its predictive policing technology, Feb 2018. URL https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd. The Verge. [Online accessed; 1/20/21].

Min Xie and Eric P. Baumer. Crime victims' decisions to call the police: Past research and new directions.

*Annual Review of Criminology*, 2(1):217–240, 2019.

Min Xie and Janet L Lauritsen. Racial context and crime reporting: A test of black's stratification hypothesis. *Journal of quantitative criminology*, 28(2):265–293, 2012.

Min Xie, Greg Pogarsky, James P Lynch, and David McDowall. Prior police contact and subsequent victim reporting: Results from the ncvs. *Justice quarterly*, 23(4):481–501, 2006.

Ke Yang. Mirror data generator package, 2023. URL `https://github.com/DataResponsibly/MirrorDataGenerator`.

Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM 2017. ACM, 2017. ISBN 9781450352826.

Sirui Yao, Yoni Halpern, Nithum Thain, Xuezhi Wang, Kang Lee, Flavien Prost, Ed H. Chi, Jilin Chen, and Alex Beutel. Measuring recommender system effects with simulated users. 2021. URL `http://arxiv.org/abs/2101.04526`.

Meike Zehlike and Carlos Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*, WWW '20, page 2849–2855. Association for Computing Machinery, Apr 2020. ISBN 9781450370233.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 Conference on Information and Knowledge Management*, CIKM '17, page 1569–1578. ACM, 2017. ISBN 9781450349185.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Lening Zhang, Steven F. Messner, and Jianhong Liu. An exploration of the determinants of reporting crime to the police in the city of tianjin, china. *Criminology*, 45(4):959–984, 2007.

Yiliang Zhang and Qi Long. Assessing fairness in the presence of missing data. In *Neurips*, 2021.

Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei Yin. Deep reinforcement learning for search, recommendation, and online advertising: a survey. *ACM SIGWEB Newsletter*, pages 1–15, 2019.

Z Zhong. A tutorial on fairness in machine learning, 2018. URL https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

Helen Zhou, Balakrishnan Sivaraman, and Zachary C. Lipton. Domain adaptation under missingness shift. *arXiv preprint, arXiv:2211.02093*, 2022.

Yan Zhou, Murat Kantarcioglu, and Chris Clifton. Improving fairness of ai systems with lossless debiasing. *arXiv preprint arXiv:2105.04534*, 2021.

# Appendix

# Appendix A

# Appendix for Chapter 2

## A.1 Supplementary figures

The following includes additional figures for Sections 2.3 and 2.4. Figures A.2, A.3, and A.5 are extended versions of Figures 2.2, 2.3, and 2.5 from the main text covering all districts of Bogotá.

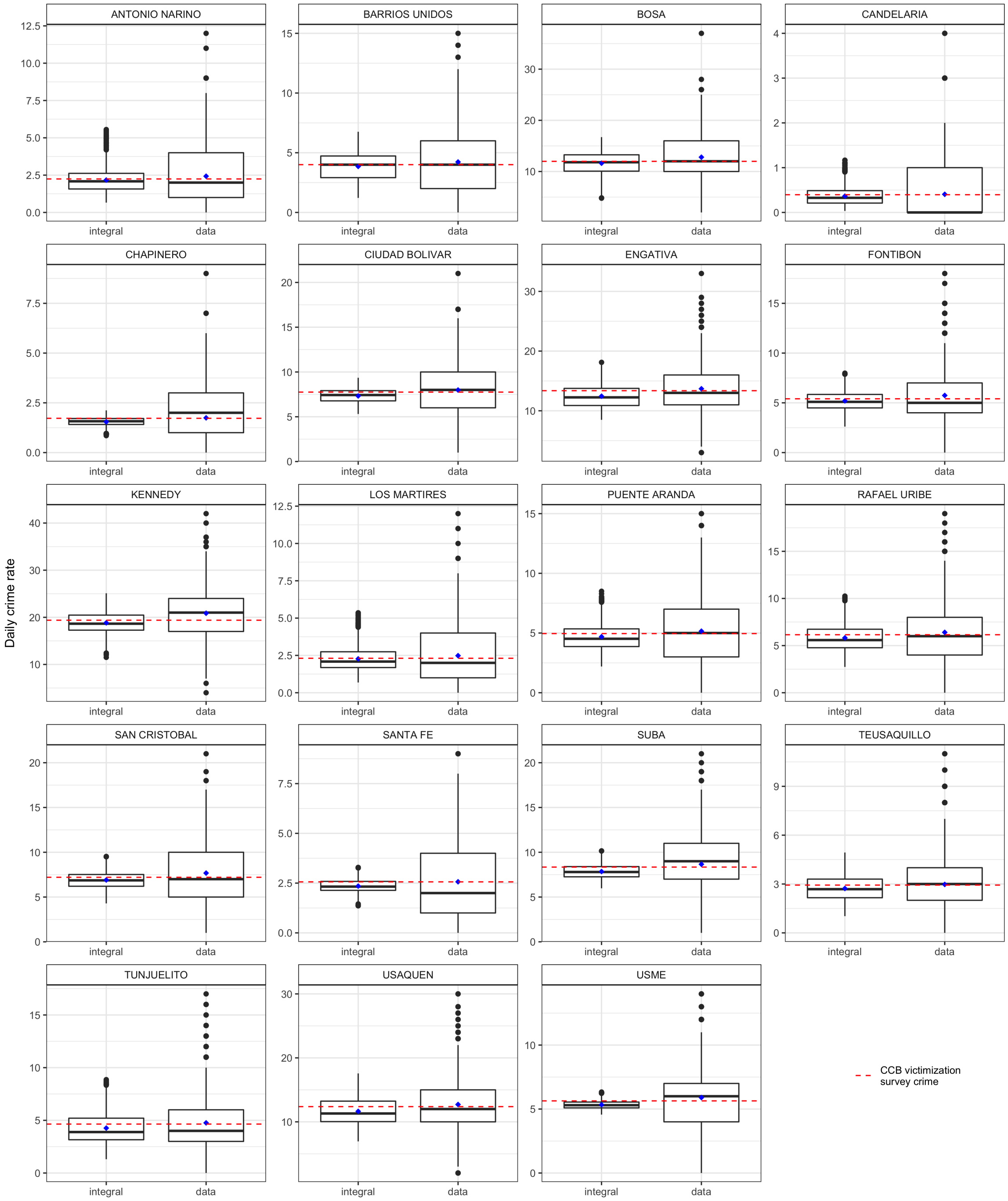


Figure A.1: District-wise sanity check of synthetic crime data over all 2,190 time steps. The average daily counts of simulated data align well with the rates obtained by integration of the data generating thinned SEPP and the desired rates implied by the CCB victimization survey.

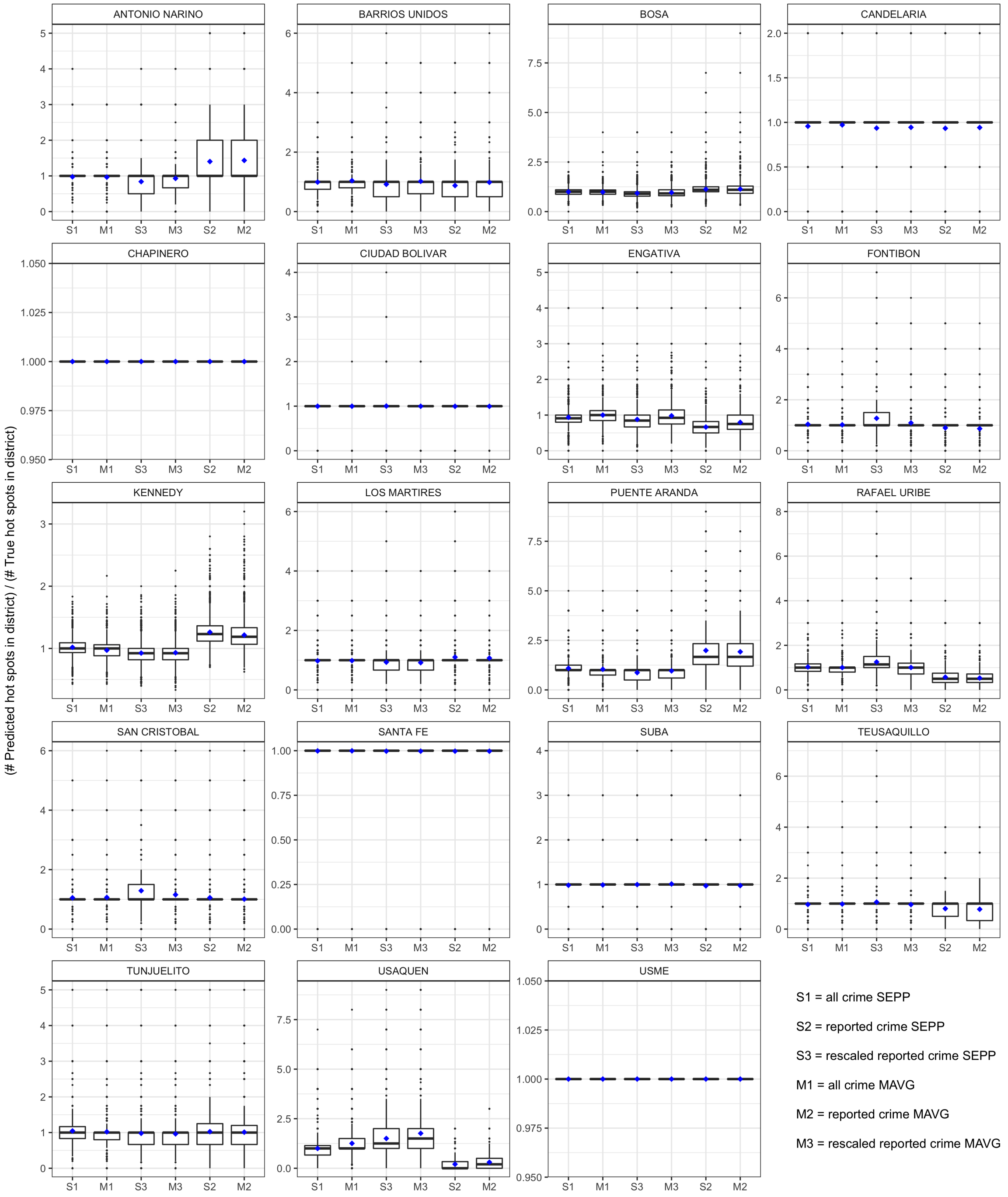


Figure A.2: Relative number of predicted crime hot spots in Bogotá districts. Each data point represents a district-specific fraction at a given evaluation time step (189 days) in a given simulation run (50 runs). A total of 50 hot spots are selected at each time step. If both the true and predicted number of hot spots is zero, we set the relative count to one. Cases for which the number of predicted hot spots is non-zero but no true hot spots are available are excluded for visualization.

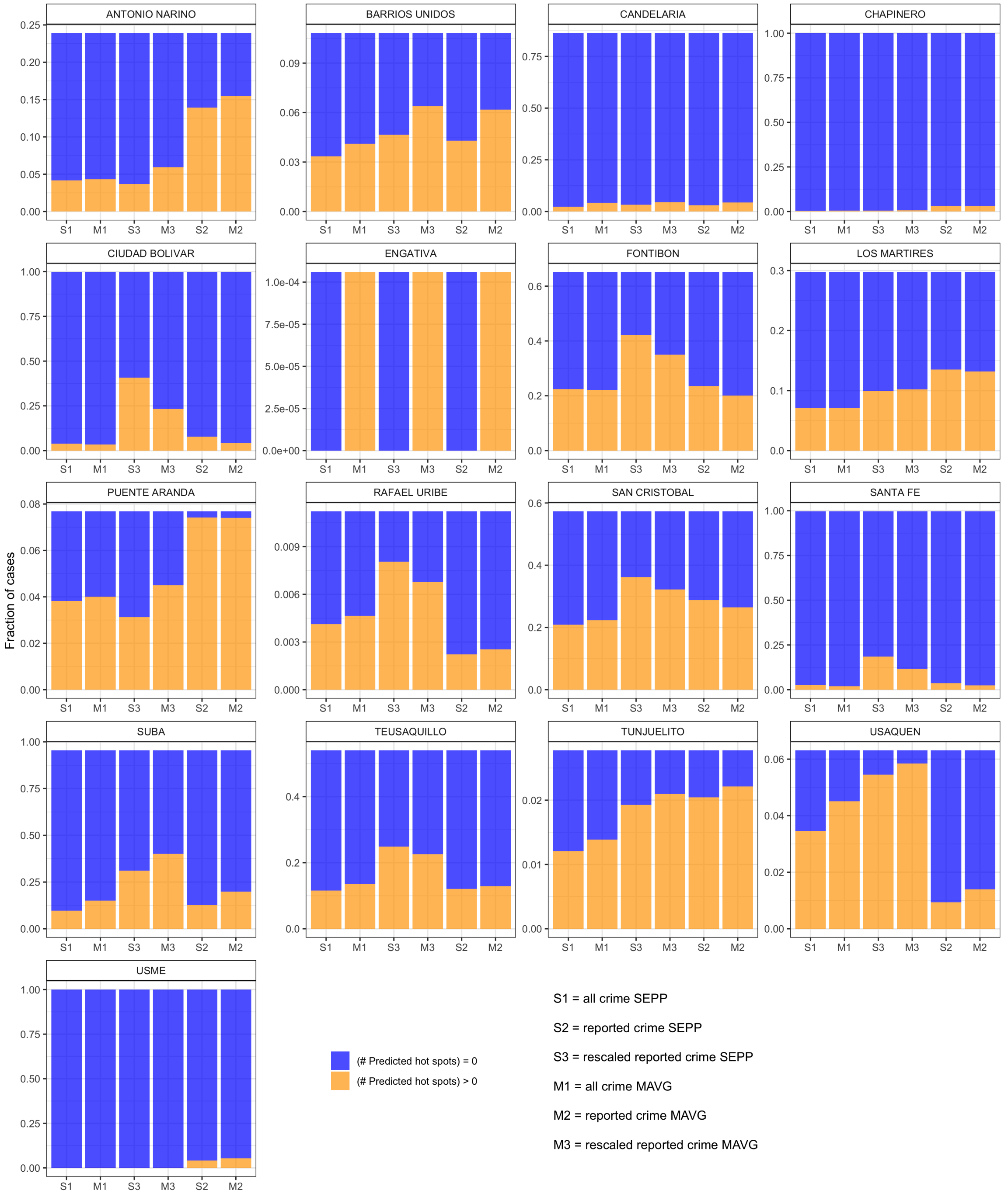


Figure A.3: Fraction of prediction time steps with no true hot spots in district separated into instances with predicted and no predicted hot spots. Ratios are computed over all evaluation time steps (189 days) and all simulation runs (50 runs) with 50 hot spots selected at each step.

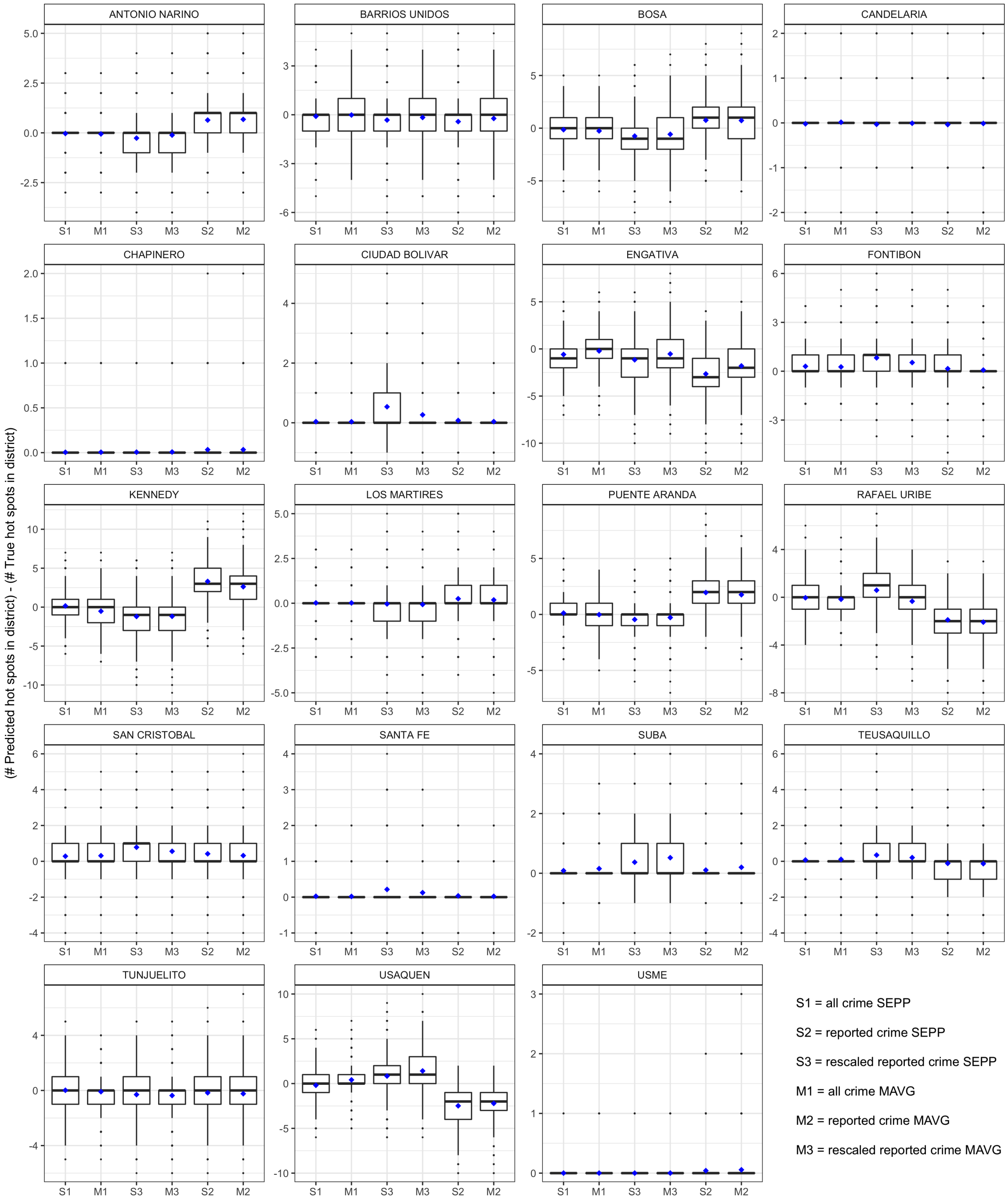


Figure A.4: Absolute number of overpredicted hot spots over all evaluation time steps (189 days) and all simulation runs (50 runs) with 50 hot spots selected at each step.

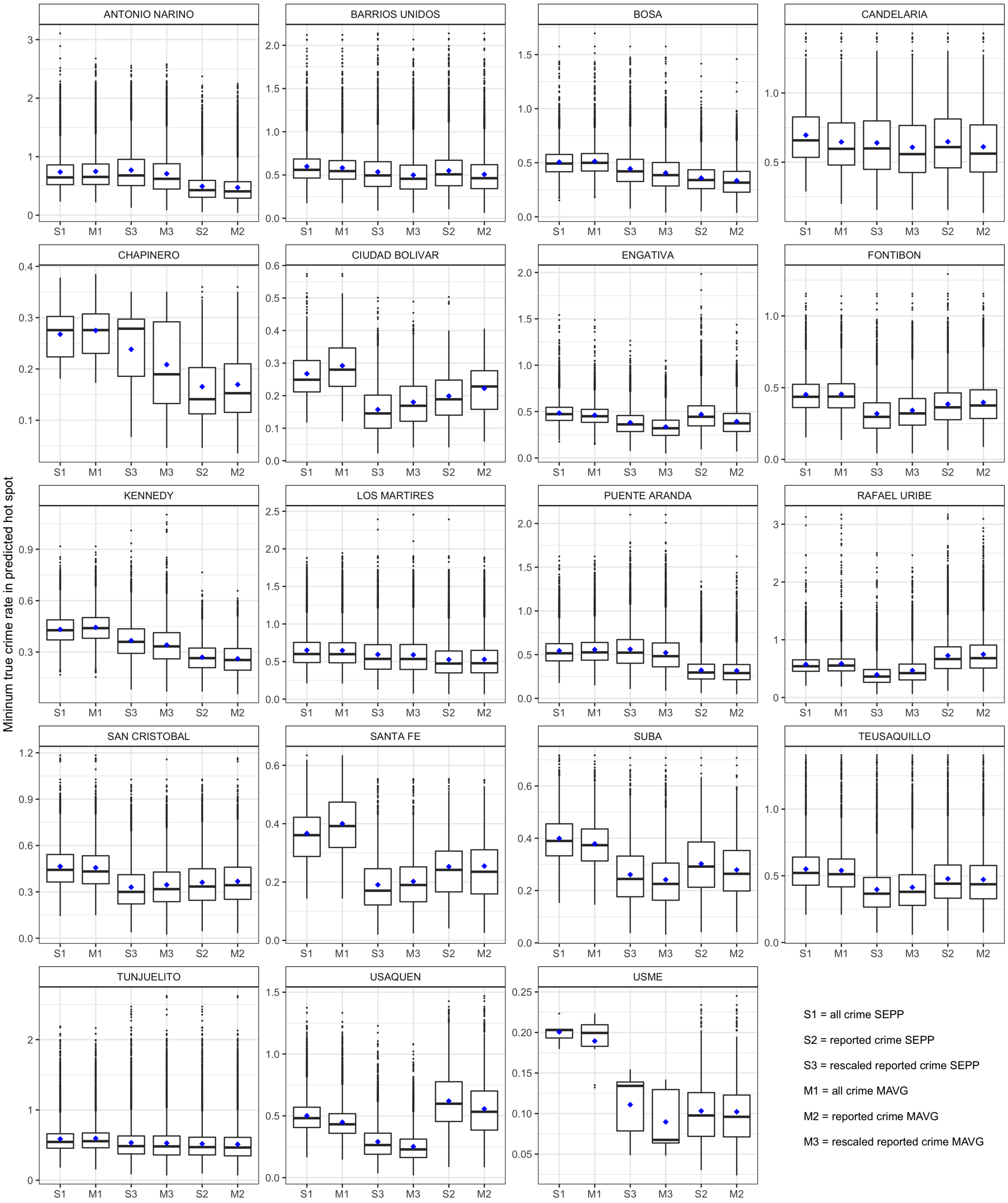


Figure A.5: True crime thresholds for hot spot selection in Bogotá districts. Each point corresponds to an evaluation time step (189 days) and a simulation run (50 runs). A total of 50 hot spots is selected at each step, and cases in which no hot spot is predicted within the district are omitted for visualization.

# Appendix B

# Appendix for Chapter 3

## B.1 Proofs

In this section, we provide the full proofs for the results in the main text.

**Lemma 3** We have $\xi \perp Z$ and $\mathbb{E}\left[\xi^2\right] = \mathbb{E}[\xi]$. Since $\mathbb{E}[\xi] \in [0,1]$, we have

$$|\hat{\beta}| = | \frac{\operatorname{Cov}[X, Z]}{\operatorname{Cov}[X, X]}\beta | = | \frac{\mathbb{E}\left[\xi Z^2\right] - \mathbb{E}[\xi Z]\,\mathbb{E}[Z]}{\mathbb{E}[\xi^2 Z^2] - \mathbb{E}[\xi Z]^2}\beta | = \frac{\mathbb{E}[\xi]\,\mathbb{V}[Z]}{\mathbb{E}[\xi]\left(\mathbb{E}[Z^2] - \mathbb{E}[\xi]\,\mathbb{E}[Z]^2\right)} | \beta | \leq | \beta | .$$

**Proposition 4** In order to derive the equations for $\hat{\beta}_i$ for $i \in [1:d]$, we start by inverting the covariance matrix

$$\Sigma_X = \begin{pmatrix} \mathbb{V}[Z_1\xi_1] & \operatorname{Cov}[Z_1\xi_1, Z_2] & \operatorname{Cov}[Z_1\xi_1, Z_3] & \cdots & \operatorname{Cov}[Z_1\xi_1, Z_d] \\ \operatorname{Cov}[Z_2, Z_1\xi_1] & \mathbb{V}[Z_2] & \operatorname{Cov}[Z_2, Z_3] & \cdots & \operatorname{Cov}[Z_2, Z_d] \\ \operatorname{Cov}[Z_3, Z_1\xi_1] & \operatorname{Cov}[Z_3, Z_2] & \mathbb{V}[Z_3] & \cdots & \operatorname{Cov}[Z_3, Z_d] \\ \vdots & & & \ddots & \vdots \\ \operatorname{Cov}[Z_d, Z_1\xi_1] & \operatorname{Cov}[Z_d, Z_2] & \cdots & & \mathbb{V}[Z_d] \end{pmatrix}.$$

For this, we separate the matrix into the blocks $A = (\Sigma_X)_{11}$, $B = ((\Sigma_X)_{1j})_{j\in[2:d]}$, $C = ((\Sigma_X)_{i1})_{i\in[2:d]}$, and $D = ((\Sigma_X)_{ij})_{i,j\in[2:d]}$. Note that $A \in \mathbb{R}^{1\times 1}$, $B = C^T \in \mathbb{R}^{1\times(d-1)}$, and $D \in \mathbb{R}^{(d-1)\times(d-1)}$. Using matrix inversion theorem, the inverse of $\Sigma_X$ can be written as

$$(\Sigma_X)^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}, \tag{B.1}$$

where $D - CA^{-1}B = D - CA^{-1}C^T$ is the Schur complement of $A$ in $\Sigma_X$. Recall that, by assumption, $\mathrm{Cov}\left[Z_i, Z_j\right] = 0$ for $i, j > 1$ with $i \neq j$ which means that $D$ is a diagonal matrix. We also note that $\mathrm{rank}(CC^T) = 1$. Denoting $g = \mathrm{trace}(-A^{-1}CC^TD^{-1})$, the inverse of the Schur complement can be written as

$$(D - CA^{-1}C^T)^{-1} = (D - A^{-1}CC^T)^{-1} = D^{-1} + \frac{1}{1+g}D^{-1}A^{-1}CC^TD^{-1}.$$

Here, $D^{-1}$ is a diagonal matrix with values $1/\mathbb{V}\left[Z_i\right]$ for $i \in [2:d]$ on the diagonal, $A^{-1} = 1/\mathbb{V}\left[Z_1\xi_1\right]$, and the diagonal of $CC^T$ is $\mathrm{Cov}\left[Z_1\xi_1, Z_i\right]^2$ for $i \in [2:d]$. It follows that

$$g = -\sum_{i=2}^{d} \rho(Z_1\xi_1, Z_i)^2$$

which corresponds to the negative of the $R^2$ between $Z_1\xi$ and $Z_2, \ldots, Z_d$. We hence write $R^2$ for $-g$ in the following.

Now we can calculate the top left block of the inverse matrix in Equation B.1 as

$$A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} = \frac{1}{\mathbb{V}\left[Z_1\xi_1\right]}\frac{1}{1-R^2}.$$

The top right bock corresponds to the row vector

$$-AB(D - CA^{-1}B)^{-1} = \left(\left(-\frac{\mathrm{Cov}\left[Z_1\xi_1, Z_i\right]}{\mathbb{V}\left[Z_i\right]\mathbb{V}\left[Z_1\xi_1\right]}\frac{1}{1-R^2}\right)_i\right)_{i\in[2:d]},$$

while the bottom left block is the same transposed. Lastly, the bottom left block of Equation B.1 can be computed as

$$(D - CA^{-1}B)^{-1} = \left( \operatorname{diag}(1/\mathbb{V}\left[Z_i\right]) + \frac{1}{1-R^2} \left( \frac{\operatorname{Cov}\left[Z_1\xi_1, Z_i\right] \operatorname{Cov}\left[Z_1\xi_1, Z_j\right]}{\mathbb{V}\left[Z_1\xi_1\right] \mathbb{V}\left[Z_i\right] \mathbb{V}\left[Z_j\right]} \right)_{ij} \right)_{i,j\in[2:d]} .$$

Inserting these values into Equation 3.1 yields the desired parameter estimates

$$\hat{\beta}_1 = \beta_1 \frac{1}{1-R^2} \sqrt{\frac{\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[X_1\right]}} \left( \rho(X_1, Z_1) - \sum_{i=2}^{d} \rho(X_1, Z_i)\rho(Z_1, Z_i) \right),$$

and

$$\hat{\beta}_k = \beta_1 \sqrt{\frac{\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_k\right]}} \left( \rho(Z_k, Z_1) - \frac{1}{1-R^2} \rho(X_1, Z_k) \left( \rho(X_1, Z_1) - \sum_{i=2}^{d} \rho(X_1, Z_i)\rho(Z_1, Z_i) \right) \right) + \beta_k.$$

**Proposition 5** The estimates from Proposition 4 simplify to

$$\hat{\beta}_1 = \frac{1}{1-R^2} \left( \frac{\mathbb{E}\left[\xi_1\right] \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]} \right) \beta_1$$

and

$$\hat{\beta}_k = \beta_1 \frac{1}{1-R^2} \left( \frac{\operatorname{Cov}\left[Z_1, Z_k\right] (\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right] \mathbb{E}\left[\xi_1\right]^2)}{\mathbb{V}\left[Z_k\right] \mathbb{V}\left[Z_1\xi_1\right]} \right) + \beta_k,$$

for $k \in [2:d]$.

For the first claim, recall that $\mathbb{V}\left[\xi\right] = \mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2$ and $\mathbb{V}\left[Z_1\xi_1\right] = \mathbb{V}\left[Z_1\right] \mathbb{E}\left[\xi_1\right]^2 + \mathbb{V}\left[\xi_1\right] \mathbb{E}\left[Z_1^2\right]$. Note that $Z \perp \xi$ allows us to rewrite

$$\begin{aligned} R^2 &= \sum_{i=2}^{d} \rho(Z_1\xi_1, Z_i)^2 \\ &= \sum_{i=2}^{d} \frac{\mathbb{E}\left[\xi_1\right]^2 \operatorname{Cov}\left[Z_1, Z_i\right]^2 \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right] \mathbb{V}\left[Z_i\right] \mathbb{V}\left[Z_1\right]} \\ &= \frac{\mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} \sum_{i=2}^{d} \rho(Z_1, Z_i)^2 \end{aligned}$$

and thus

$$\begin{aligned}
&\frac{1}{1-R^2}\left(\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)\\
=&\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right] - R^2\mathbb{V}\left[Z_1\xi_1\right]}{(1-R^2)\mathbb{V}\left[Z_1\xi_1\right]\mathbb{E}\left[\xi_1\right]}\\
=&\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right] - \frac{\mathbb{E}[\xi_1]^2\mathbb{V}[Z_1]}{\mathbb{V}[Z_1\xi_1]}\sum_{i=2}^d \rho(Z_1,Z_i)^2\mathbb{V}\left[Z_1\xi_1\right]}{(1-\frac{\mathbb{E}[\xi_1]^2\mathbb{V}[Z_1]}{\mathbb{V}[Z_1\xi_1]}\sum_{i=2}^d \rho(Z_1,Z_i)^2)\mathbb{V}\left[Z_1\xi_1\right]\mathbb{E}\left[\xi_1\right]}\\
=&\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2\mathbb{V}\left[Z_1\right]\sum_{i=2}^d \rho(Z_1,Z_i)^2}{(\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2 + (\mathbb{E}\left[\xi_1\right]-\mathbb{E}\left[\xi_1\right]^2)\mathbb{E}\left[Z_1^2\right])\mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2\mathbb{V}\left[Z_1\right]\sum_{i=2}^d \rho(Z_1,Z_i)^2\mathbb{E}\left[\xi_1\right]}\\
=&\frac{\mathbb{V}\left[Z_1\right](1-\sum_{i=2}^d \rho(Z_1,Z_i)^2)}{\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right](1-\sum_{i=2}^d \rho(Z_1,Z_i)^2) + (1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]},
\end{aligned}$$

which is positive as long as $Z_1$ is not a linear combination of other features which was explicitly excluded from consideration. The claim follows.

For the second claim, we show that

$$\begin{aligned}
&\frac{\mathbb{V}\left[Z_1\right](1-\sum_{i=2}^d \rho(Z_1,Z_i)^2)}{\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right](1-\sum_{i=2}^d \rho(Z_1,Z_i)^2) + (1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]} < 1\\
\Leftrightarrow&(1-\sum_{i=2}^d \rho(Z_1,Z_i)^2)\mathbb{V}\left[Z_1\right](1-\mathbb{E}\left[\xi_1\right]) < (1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]\\
\Leftrightarrow&(1-\sum_{i=2}^d \rho(Z_1,Z_i)^2)(\mathbb{E}\left[Z_1^2\right]-\mathbb{E}\left[Z_1\right]^2) < \mathbb{E}\left[Z_1^2\right]\\
\Leftrightarrow&(1-\sum_{i=2}^d \rho(Z_1,Z_i)^2)\left(1-\frac{\mathbb{E}\left[Z_1\right]^2}{\mathbb{E}\left[Z_1\right]^2}\right) < 1.
\end{aligned}$$

Since $Z_1$ is not a linear combination of other features, we know that $1-\sum_{i=2}^d \rho(Z_1,Z_i)^2 \in (0,1]$ and this inequatily is always true. The claim follows with the first part of the proposition.

For the third claim, recall that $\mathbb{E}\left[\xi_1\right] = rm_1^1 + (1-r)m_1^0$. Assume we have two sets of parameters $(m_1^0, m_1^1)$ and $(m_1^{0'}, m_1^{1'})$. If $m_1^0 < m_1^{0'}$ and $m_1^1 \leq m_1^{1'}$, we have

$$\mathbb{E}\left[\xi\right] = rm_1^1 + (1-r)m_1^0 < rm_1^{1'} + (1-r)m_1^{0'} = \mathbb{E}\left[\xi'\right].$$

The same holds true if $m_1^0 \leq m_1^{0'}$ and $m_1^1 < m_1^{1'}$ which shows that the expected share of observed features $\mathbb{E}\left[\xi_1\right]$ is decreasing if and only if we are increasing missingness in either (or both) of the groups while leaving everything else fixed. Thus, instead of changes in $m_1^g$, we argue directly about changes in $\mathbb{E}\left[\xi_1\right]$ in the following.

Denote $S^2 = \sum_{i=2}^d \rho(Z_1, Z_i)^2$ and consider the function

$$\begin{aligned} &f : (0,1] \to \mathbb{R} \\ &\mathbb{E}\left[\xi_1\right] = x \mapsto \frac{\mathbb{V}\left[Z_1\right](1-S^2)}{\mathbb{V}\left[Z_1\right]x(1-S^2) + (1-x)\mathbb{E}\left[Z_1^2\right]}. \end{aligned}$$

We show that $f$ is monotonically increasing from which the claim follows directly. It holds that

$$\frac{d}{dx}f(x) = \frac{-\mathbb{V}\left[Z_1\right](1-S^2)(\mathbb{V}\left[Z_1\right](1-S^2) - \mathbb{E}\left[Z_1^2\right])}{\left(\mathbb{V}\left[Z_1\right]x(1-S^2) + (1-x)\mathbb{E}\left[Z_1^2\right]\right)^2}.$$

Since $1 - S^2 \in [0,1)$ and $\mathbb{V}\left[Z_1\right] > 0$, the numerator is positive iff

$$\begin{aligned} &-\mathbb{V}\left[Z_1\right](1-S^2)(\mathbb{V}\left[Z_1\right](1-S^2) - \mathbb{E}\left[Z_1^2\right]) > 0 \\ \Leftrightarrow &\mathbb{V}\left[Z_1\right](1-S^2) < \mathbb{E}\left[Z_1^2\right] \\ \Leftrightarrow &\left(1 - \frac{\mathbb{E}\left[Z_1\right]^2}{\mathbb{E}\left[Z_1^2\right]}\right)(1-S^2) < 1, \end{aligned}$$

which is a true statement. We conclude that $f$ is monotonically increasing in $x$ and the claim follows.

**Proposition 6** Given a $k \in [2:d]$, we know that

$$\hat{\beta}_k = \beta_1 \frac{1}{1-R^2}\left(\frac{\text{Cov}\left[Z_1, Z_k\right](\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2)}{\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\xi_1\right]}\right) + \beta_k.$$

The first claim is obvious from this expression.

For the second claim, we follow similar steps as for the third claim in the proof of Proposition 5. We know from the previous proof that

$$R^2 = \frac{\mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1 \xi_1\right]} \sum_{i=2}^{d} \rho(Z_1, Z_i)^2$$

and thus, denoting $S^2 = \sum_{i=2}^{d} \rho(Z_1, Z_i)^2$,

$$\begin{aligned}
&\frac{1}{1-R^2}\left(\frac{\text{Cov}\left[Z_1, Z_k\right](\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_2\right]\mathbb{E}\left[\xi\right]^2)}{\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\xi_1\right]}\right)\\
=&\frac{\text{Cov}\left[Z_1, Z_k\right]\mathbb{V}\left[\xi_1\right]\mathbb{E}\left[Z_1^2\right]}{(1-\frac{\mathbb{E}[\xi_1]^2\mathbb{V}[Z_1]}{\mathbb{V}[Z_1\xi_1]}S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\xi_1\right]}\\
=&\frac{\text{Cov}\left[Z_1, Z_k\right]\mathbb{V}\left[\xi_1\right]\mathbb{E}\left[Z_1^2\right]}{\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2\mathbb{V}\left[Z_1\right]S^2\mathbb{V}\left[Z_k\right]}\\
=&\frac{\text{Cov}\left[Z_1, Z_k\right](\mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2)\mathbb{E}\left[Z_1^2\right]}{\mathbb{V}\left[Z_k\right](\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2 + (\mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2)\mathbb{E}\left[Z_1^2\right]) - \mathbb{E}\left[\xi_1\right]^2\mathbb{V}\left[Z_1\right]S^2\mathbb{V}\left[Z_k\right]}\\
=&\frac{\text{Cov}\left[Z_1, Z_k\right](1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]}{(1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right] + \mathbb{V}\left[Z_k\right](1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]},
\end{aligned}$$

since $\mathbb{V}\left[Z_1\xi_1\right] = \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2 + \mathbb{V}\left[\xi_1\right]\mathbb{E}\left[Z_1^2\right]$ and $\mathbb{V}\left[\xi_1\right] = \mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2$.

Now, consider the function

$$\begin{aligned}
&g : (0, 1] \to \mathbb{R}\\
\mathbb{E}\left[\xi_1\right] = x \mapsto &\frac{\text{Cov}\left[Z_1, Z_k\right](1-x)\mathbb{E}\left[Z_1^2\right]}{(1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right]x + \mathbb{V}\left[Z_k\right](1-x)\mathbb{E}\left[Z_1^2\right]}.
\end{aligned}$$

We compute

$$\begin{aligned}
\frac{d}{dx}g(x) =&\frac{-\text{Cov}\left[Z_1, Z_k\right]\mathbb{E}\left[Z_1^2\right]((1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right]x + \mathbb{V}\left[Z_k\right](1-x)\mathbb{E}\left[Z_1^2\right])}{\left((1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right]x + \mathbb{V}\left[Z_k\right](1-x)\mathbb{E}\left[Z_1^2\right]\right)^2}\\
&-\frac{(1-x)\text{Cov}\left[Z_1, Z_k\right]\mathbb{E}\left[Z_1^2\right]((1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right] - \mathbb{V}\left[Z_k\right]\mathbb{E}\left[Z_1^2\right])}{\left((1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right]x + \mathbb{V}\left[Z_k\right](1-x)\mathbb{E}\left[Z_1^2\right]\right)^2}.
\end{aligned}$$

Further,

$$\begin{aligned}
&\frac{d}{dx}g(x) > 0 \\
\Leftrightarrow &-\mathrm{Cov}\left[Z_1, Z_k\right]\mathbb{E}\left[Z_1^2\right]((1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right]x + \mathbb{V}\left[Z_k\right](1-x)\mathbb{E}\left[Z_1^2\right]) \\
&> (1-x)\mathrm{Cov}\left[Z_1, Z_k\right]\mathbb{E}\left[Z_1^2\right]((1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right] - \mathbb{V}\left[Z_k\right]\mathbb{E}\left[Z_1^2\right]) \\
\Leftrightarrow &-\mathrm{Cov}\left[Z_1, Z_k\right](1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right]x > (1-x)\mathrm{Cov}\left[Z_1, Z_k\right](1-S^2)\mathbb{V}\left[Z_k\right]\mathbb{V}\left[Z_1\right] \\
\Leftrightarrow &0 > \mathrm{Cov}\left[Z_1, Z_k\right].
\end{aligned}$$

This shows that factor determining the influence of $\beta_1$ on $\hat{\beta}_k$ is increasing with decreasing missingness if $\mathrm{Cov}\left[Z_1, Z_k\right] < 0$ and decreasing with decreasing missingness otherwise. The claim follows.

**Proposition 7** Predictions are obtained from the model $\hat{Y} = \hat{\alpha} + \hat{\beta}^T X$. Since $Z \sim \mathcal{N}(\mu, \Sigma)$ is jointly Gaussian, we know that

$$\hat{\beta}^T Z \sim \mathcal{N}\left(\hat{\beta}^T\mu, \hat{\beta}^T\Sigma\hat{\beta}\right) = \mathcal{N}\left(\sum_{i=1}^{d}\hat{\beta}_i\mu_i, \sum_{i=1}^{d}\hat{\beta}_i^2\sigma_i^2 + \sum_{i=1}^{d}\sum_{j=i+1}^{d}2\hat{\beta}_i\hat{\beta}_j\mathrm{Cov}\left[Z_i, Z_j\right]\right)$$

and

$$\hat{\beta}_{[2:d]}^T Z_{[2:d]} \sim \mathcal{N}\left(\hat{\beta}_{[2:d]}^T\mu_{[2:d]}, \hat{\beta}_{[2:d]}^T\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}\right) = \mathcal{N}\left(\sum_{i=2}^{d}\hat{\beta}_i\mu_i, \sum_{i=2}^{d}\hat{\beta}_i^2\sigma_i^2 + \sum_{i=2}^{d}\sum_{j=i+1}^{d}2\hat{\beta}_i\hat{\beta}_j\mathrm{Cov}\left[Z_i, Z_j\right]\right)$$

where $\sigma_i^2 = \mathbb{V}\left[Z_i\right]$ for $i \in [1:d]$.

The cdf of predictions $\hat{Y}$ in group $g \in \{0, 1\}$ can be written as

$$\begin{aligned}
F_{\hat{Y}|G=g}(x) =& P\left(\hat{\beta}_1 Z_1 \xi_1^g + \hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq x - \hat{\alpha}\right) \\
=& (1-m_1^g)P\left(\hat{\beta}_{[2:d]}^T Z_{[2:d]} \leq x - \hat{\alpha}\right) + m_1^g P\left(\hat{\beta}^T Z \leq x - \hat{\alpha}\right).
\end{aligned}$$

Let $C \in [0, 1]$ and denote $\tilde{y} = F_{\hat{Y}}^{-1}(1-C)$. Without loss of generality, assume that $m_1^0 < m_1^1$. If $m_1^0 = m_1^1$ the selection rate disparity is 0, if $m_1^0 > m_1^1$ the following calculation can easily be adjusted. The inequality $m_1^0 < m_1^1$ means that group 0 has the same or more expected missingness in feature $Z_1$

than group 1. Group 0 is over-selected at threshold $C$ according to Definition 2 if and only if

$$\begin{aligned}
&1 - F_{\hat{Y}|G=1}(\tilde{y}) < 1 - F_{\hat{Y}|G=0}(\tilde{y}) \\
\Leftrightarrow &(1 - m_1^1)P\left(\hat{\beta}^T_{[2:d]} Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right) + m_1^1 P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) > (1 - m_1^0)P\left(\hat{\beta}^T_{[2:d]} Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right) + m_1^0 P\left(\hat{\beta}^T Z \leq \tilde{y} -\right. \\
\Leftrightarrow &(m_1^1 - m_1^0)P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) > (m_1^1 - m_1^0)P\left(\hat{\beta}^T_{[2:d]} Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right) \\
\Leftrightarrow &P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) > P\left(\hat{\beta}^T_{[2:d]} Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right).
\end{aligned}$$

Expanding on this in the jointly Gaussian case, we can see that

$$\begin{aligned}
&P\left(\hat{\beta}^T Z \leq \tilde{y} - \hat{\alpha}\right) > P\left(\hat{\beta}^T_{[2:d]} Z_{[2:d]} \leq \tilde{y} - \hat{\alpha}\right) \\
\Leftrightarrow &\Phi\left(\frac{\tilde{y} - \hat{\alpha} - \hat{\beta}^T \mu}{\sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}\right) > \Phi\left(\frac{\tilde{y} - \hat{\alpha} - \hat{\beta}^T_{[2:d]}\mu_{[2:d]}}{\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}}}\right) \\
\Leftrightarrow &\left(\tilde{y} - \hat{\alpha} - \hat{\beta}^T \mu\right)\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}} > \left(\tilde{y} - \hat{\alpha} - \hat{\beta}^T_{[2:d]}\mu_{[2:d]}\right)\sqrt{\hat{\beta}^T \Sigma \hat{\beta}} \\
\Leftrightarrow &\tilde{y}\left(\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}}\right) > \left(\hat{\alpha} + \hat{\beta}^T_{[2:d]}\mu_{[2:d]}\right)\left(\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}}\right) + \left(\hat{\beta}_1 \mu_1\right)\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:}}
\end{aligned}$$

Here, $\Phi$ is the standard normal cdf. If

$$\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}} > 0,$$

group 0 is over-selected if and only if $C$ implies a threshold $\tilde{y}$ with

$$\tilde{y} > \left(\hat{\alpha} + \hat{\beta}^T_{[2:d]}\mu_{[2:d]}\right) + \left(\hat{\beta}_1 \mu_1\right)\frac{\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}}}{\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}.$$

If

$$\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}} < 0,$$

group 0 is over-selected if and only if $C$ implies a threshold $\tilde{y}$ with

$$\tilde{y} < \left(\hat{\alpha} + \hat{\beta}^T_{[2:d]}\mu_{[2:d]}\right) + \left(\hat{\beta}_1 \mu_1\right)\frac{\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}}}{\sqrt{\hat{\beta}^T_{[2:d]}\Sigma_{[2:d,2:d]}\hat{\beta}_{[2:d]}} - \sqrt{\hat{\beta}^T \Sigma \hat{\beta}}}.$$

The proposition follows.

**Corollary 8** We combine Proposition 7 with the parameter estimates given in the proof of Proposition 5. For a high threshold $\tilde{y}$, the group with more missingness is over-selected if

$$\begin{aligned}
&\mathbb{V}\left[\hat{\beta}_{[2:d]}^T Z_{[2:d]}\right] > \mathbb{V}\left[\hat{\beta} Z\right] \\
\Leftrightarrow &\sum_{i=2}^{d} \hat{\beta}_i^2 \sigma_i^2 + \sum_{i=2}^{d} \sum_{j=i+1}^{d} 2\hat{\beta}_i \hat{\beta}_j \text{Cov}\left[Z_i, Z_j\right] > \sum_{i=1}^{d} \hat{\beta}_i^2 \sigma_i^2 + \sum_{i=1}^{d} \sum_{j=i+1}^{d} 2\hat{\beta}_i \hat{\beta}_j \text{Cov}\left[Z_i, Z_j\right] \\
\Leftrightarrow &\hat{\beta}_1^2 \sigma_1^2 + \sum_{j=2}^{d} 2\hat{\beta}_1 \hat{\beta}_j \text{Cov}\left[Z_1, Z_j\right] < 0.
\end{aligned}$$

Recall that

$$R^2 = \frac{\mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1 \xi_1\right]} \sum_{i=2}^{d} \rho(Z_1, Z_i)^2$$

and denote $S^2 = \sum_{i=2}^{d} \rho(Z_1, Z_i)^2$. Then

$$\frac{1}{1 - R^2} = \frac{\mathbb{V}\left[Z_1 \xi_1\right]}{\mathbb{V}\left[Z_1 \xi_1\right] - \mathbb{E}\left[\xi_1\right]^2 \mathbb{V}\left[Z_1\right] S^2}.$$

Using Proposition 5 and inserting the parameter estimates gives

$$\begin{aligned}
&\hat{\beta}_1^2\sigma_1^2 + \sum_{j=2}^{d} 2\hat{\beta}_1\hat{\beta}_j \mathrm{Cov}\left[Z_1, Z_j\right] < 0 \\
\Leftrightarrow &\left(\frac{1}{1-R^2}\right)^2 \left(\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)^2 \beta_1^2 \mathbb{V}\left[Z_1\right] \\
&+ \frac{1}{1-R^2}\left(\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)\beta_1 \sum_{j=2}^{d} 2\left(\beta_1 \frac{1}{1-R^2}\left(\frac{\mathrm{Cov}\left[Z_1, Z_j\right](\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2)}{\mathbb{V}\left[Z_j\right]\mathbb{V}\left[Z_1\xi_1\right]}\right) + \beta_j\right) \mathrm{Cov}\,[Z_1 \\
\Leftrightarrow &\left(\frac{1}{1-R^2}\right)\left(\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)\beta_1^2 \mathbb{V}\left[Z_1\right] \\
&+ 2\frac{1}{1-R^2}\beta_1^2 \sum_{j=2}^{d}\left(\frac{\mathrm{Cov}\left[Z_1, Z_j\right](\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2)}{\mathbb{V}\left[Z_j\right]\mathbb{V}\left[Z_1\xi_1\right]}\right)\mathrm{Cov}\left[Z_1, Z_j\right] + 2\beta_1 \sum_{j=2}^{d} \beta_j \mathrm{Cov}\left[Z_1, Z_j\right] < 0 \\
\Leftrightarrow &\frac{1}{1-R^2}\left(\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} - \frac{R^2}{\mathbb{E}\left[\xi_1\right]}\right)\beta_1^2 \mathbb{V}\left[Z_1\right] \\
&+ 2\beta_1^2 \frac{1}{1-R^2}\frac{(\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2)\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right]} \sum_{j=2}^{d}\left(\frac{\mathrm{Cov}\left[Z_1, Z_j\right]^2}{\mathbb{V}\left[Z_j\right]\mathbb{V}\left[Z_1\right]}\right) + 2\beta_1 \sum_{j=2}^{d} \beta_j \mathrm{Cov}\left[Z_1, Z_j\right] < 0 \\
\Leftrightarrow &\frac{\mathbb{E}\left[\xi_1\right]\mathbb{V}\left[Z_1\right]^2}{\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2\mathbb{V}\left[Z_1\right]S^2}(1-S^2)\beta_1^2 + 2\beta_1^2 \frac{(\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2)S^2\mathbb{V}\left[Z_1\right]}{\mathbb{V}\left[Z_1\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2\mathbb{V}\left[Z_1\right]S^2} + 2\beta_1 \sum_{j=2}^{d} \beta_j \mathrm{Cov}\left[Z_1, Z_j\right] < 0 \\
\Leftrightarrow &\beta_1^2 \mathbb{V}\left[Z_1\right] \frac{\mathbb{V}\left[Z_1\right](1-S^2) + 2(1-\mathbb{E}\left[\xi\right])\mathbb{E}\left[Z_1^2\right]S^2}{\mathbb{V}\left[Z_1\right](1-S^2)\mathbb{E}\left[\xi_1\right] + (1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]} + 2\beta_1 \sum_{j=2}^{d} \beta_j \mathrm{Cov}\left[Z_1, Z_j\right] < 0
\end{aligned}$$

where we used that $\mathbb{V}\left[Z_1\xi_1\right] = \mathbb{V}\left[Z_1\right]\mathbb{E}\left[\xi_1\right]^2 + \mathbb{V}\left[\xi_1\right]\mathbb{E}\left[Z_1^2\right]$ and $\mathbb{V}\left[\xi_1\right] = \mathbb{E}\left[\xi_1\right] - \mathbb{E}\left[\xi_1\right]^2$. Note that the first term on the left side is always positive. Thus the inequality is fulfilled if and only if

$$\mathrm{sign}\left(\beta_1 \sum_{j=2}^{d} \beta_j \mathrm{Cov}\left[Z_1, Z_j\right]\right) = -1$$

and

$$
\begin{aligned}
&2\beta_1 \sum_{j=2}^{d} \beta_j \text{Cov}\left[Z_1, Z_j\right] < -\beta_1^2 \mathbb{V}\left[Z_1\right] \frac{\mathbb{V}\left[Z_1\right](1-S^2) + 2(1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right] S^2}{\mathbb{V}\left[Z_1\right](1-S^2)\mathbb{E}\left[\xi_1\right] + (1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]} \\
\Leftrightarrow &\frac{1}{\beta_1} \sum_{j=2}^{d} \beta_j \text{Cov}\left[Z_1, Z_j\right] < -\frac{\mathbb{V}\left[Z_1\right]^2(1-S^2) + 2(1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]\mathbb{V}\left[Z_1\right] S^2}{2\mathbb{V}\left[Z_1\right](1-S^2)\mathbb{E}\left[\xi_1\right] + 2(1-\mathbb{E}\left[\xi_1\right])\mathbb{E}\left[Z_1^2\right]} \\
\Leftrightarrow &\frac{1}{\beta_1} \sum_{j=2}^{d} \beta_j \text{Cov}\left[Z_1, Z_j\right] < -\frac{\mathbb{V}\left[Z_1\right]^2(1-S^2) + 2(1-\mathbb{E}\left[\xi_1\right])(\mathbb{V}\left[Z_1\right] + \mathbb{E}\left[Z_1\right]^2)\mathbb{V}\left[Z_1\right] S^2}{2\mathbb{V}\left[Z_1\right](1-S^2)\mathbb{E}\left[\xi_1\right] + 2(1-\mathbb{E}\left[\xi_1\right])(\mathbb{V}\left[Z_1\right] + \mathbb{E}\left[Z_1\right]^2)}.
\end{aligned}
$$

Since we know that the fraction on the right side is always positive, this can be rewritten as presented in the corollary. If the inequality is not fulfilled, the group with more missingness is under-selected at a high threshold.

**Lemma 9** In the setting of the Lemma, we can write

$$
\begin{aligned}
\mathbb{E}_\xi\left[l(f, X, y)\right] &= \mathbb{E}_\xi\left[\frac{1}{m} l(f, X, y) - \frac{1-m}{m} l(f, [0, X_{[2:d]}]^T, y)\right] \\
&= \frac{1}{m}\mathbb{E}_\xi\left[l(f, X, y)\right] - \frac{1-m}{m} l(f, [0, z_{[2:d]}]^T, y) \\
&= \frac{1}{m}\left(P(\xi_1 = 1) l(f, z, y) + P(\xi_1 = 0) l(f, [0, z_{[2:d]}]^T, y)\right) - \frac{1-m}{m} l(f, [0, z_{[2:d]}]^T, y) \\
&= l(f, z, y) + \frac{1-m}{m} l(f, [0, z_{[2:d]}]^T, y) - \frac{1-m}{m} l(f, [0, z_{[2:d]}]^T, y) \\
&= l(f, z, y).
\end{aligned}
$$

Here, the first equality holds since only the first feature has under-reporting and the second equality holds because $Z \perp G$ which implies $Z \perp \xi$.

**Lemma 10** Follows the same as Lemma 9. Instead of under-reporting completely at random, the under-reporting is completely at random within group $g$.

**Lemma 11** Since $f$ is linear and under-repoting only occurs in the first feature, the expected prediction error for an imputation value $x'_1$ can be written as

$$\begin{aligned}
R(f) &= \mathbb{E}_X\left[(f(X') - Y)^2\right] \\
&= \beta_1^2 \mathbb{E}_X\left[(X'_1 - Z_1)^2\right] \\
&= \beta_1^2 P(X_1 = 0)\mathbb{E}_X\left[(X'_1 - Z_1)^2 \mid X_1 = 0\right] + \beta_1^2 P(X_1 \neq 0)\mathbb{E}_X\left[(X'_1 - Z_1)^2 \mid X_1 \neq 0\right] \\
&= \beta_1^2 P(X_1 = 0)\mathbb{E}_Z\left[(x'_1 - Z_1)^2 \mid X_1 = 0\right] + \beta_1^2 P(X_1 \neq 0)\mathbb{E}_Z\left[(Z_1 - Z_1)^2 \mid X_1 \neq 0\right] \\
&= \beta_1^2 P(X_1 = 0)\left(x_1'^2 - 2x'_1\mathbb{E}\left[Z_1 \mid X_1 = 0\right] + \mathbb{E}\left[Z_1^2 \mid X_1 = 0\right]\right).
\end{aligned}$$

Then

$$\begin{aligned}
\frac{dR(f)}{dx'_1} &= \beta_1^2 P(X_1 = 0)(2x'_1 - 2\mathbb{E}\left[Z_1 \mid X_1 = 0\right]) \overset{!}{=} 0 \\
\Leftrightarrow x'_1 &= \mathbb{E}\left[Z_1 \mid X_1 = 0\right].
\end{aligned}$$

We implicitly assume that $\beta_1 \neq 0$ and the probability of 0-entries is positive.

**Lemma 12** Recall that $G \sim \text{Bern}(r)$. Similar to the proof of Lemma 11, the expected prediction error can be written as

$$\begin{aligned}
R(f) &= \mathbb{E}_X\left[(f(X') - Y)^2\right] \\
&= \beta_1^2 P(X_1 = 0)\mathbb{E}_Z\left[(X'_1 - Z_1)^2 \mid X_1 = 0\right] \\
&= \beta_1^2 P(X_1 = 0)\left(r\mathbb{E}_Z\left[(X'_1 - Z_1)^2 \mid X_1 = 0, G = 1\right] + (1-r)\mathbb{E}_Z\left[(X'_1 - Z_1)^2 \mid X_1 = 0, G = 0\right]\right) \\
&= \beta_1^2 P(X_1 = 0)\left(r\mathbb{E}_Z\left[(x_1'^1 - Z_1)^2 \mid X_1 = 0, G = 1\right] + (1-r)\mathbb{E}_Z\left[(x_1'^0 - Z_1)^2 \mid X_1 = 0, G = 0\right]\right).
\end{aligned}$$

This prediction error is minimal when

$$\begin{aligned}
\frac{dR(f)}{dx_1'^g} &= \beta_1^2 P(X_1 = 0)P(G = g)(2x_1'^g - 2\mathbb{E}\left[Z_1 \mid X_1 = 0, G = g\right]) \overset{!}{=} 0 \\
\Leftrightarrow x_1'^g &= \mathbb{E}\left[Z_1 \mid X_1 = 0, G = g\right]
\end{aligned}$$

for $g \in \{0, 1\}$.

## B.2 Connection between Proposition 4 and omitted variable bias

Econometrics literature uses the term omitted variable bias to refer to the model estimation bias that is introduced when omitting an independent variable that influences both other independent variables and the dependent outcome (Angrist and Pischke, 2008). In the setting of Proposition 4, omitting the first feature entirely corresponds to a setting in which all feature entries are under-reported, i.e. default to 0. The $k$-th parameter estimate in this case can be written as

$$\hat{\beta}_k = \beta_1 \frac{\text{Cov}\left[Z_k, Z_1\right]}{\mathbb{V}\left[Z_k\right]} + \beta_k$$

which is known as omitted variable bias formula (Angrist and Pischke, 2008). Here,

$$\gamma_{Z_1,Z_k} = \text{Cov}\left[Z_1, Z_k\right] / \mathbb{V}\left[Z_k\right]$$

corresponds to the population regression coefficient of a linear regression of $Z_1$ on $Z_k$ which can be written as

$$Z_1 = \alpha_{Z_1,Z_k} + \gamma_{Z_1,Z_k} Z_k,$$

where $\alpha_{Z_1,Z_k}$ is an intercept. Omitting $Z_1$ from the regression induces a confounding relationship where the effects of $Z_1$ on $Z_k$ become intertwined. Instead of isolating the effect of $Z_k$ on $Y$, $\hat{\beta}_k$ also includes a partial effect of $Z_1$ on $Y$. This effect is scaled by $\gamma_{Z_1,Z_k}$ to account for the linear relationship between $Z_1$ and $Z_k$.

In the setting of this paper, we are interested in cases in which some but not necessarily all of the feature entries are missing. Maintaining the same notation as before, $\hat{\beta}_1$ from Equation 3.2 in this general case can be written as

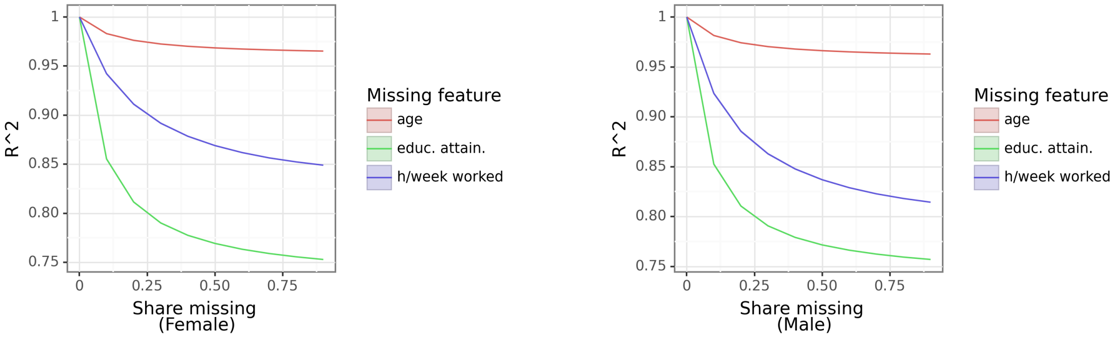
$$\begin{aligned}\hat{\beta_1} &= \beta_1 \frac{1}{1-R^2}\left(\frac{\text{Cov}\left[X_1, Z_1\right]}{\mathbb{V}\left[X_1\right]} - \sum_{i=2}^{d} \frac{\text{Cov}\left[X_1, Z_i\right] \text{Cov}\left[Z_1, Z_i\right]}{\mathbb{V}\left[X_1\right] \mathbb{V}\left[Z_i\right]}\right) \\ &= \beta_1 \left(\frac{\gamma_{Z_1,X_1} - \sum_{i=2}^{d} \gamma_{Z_i,X_1}\gamma_{Z_1,Z_i}}{1-R^2}\right).\end{aligned}$$

Here, the numerator of the biasing factor reflects how much information about $Z_1$ remains encoded in $X_1$ without drawing on associations through the other features $Z_2, \ldots, Z_d$ (i.e, arrows of the form $X_1 \to Z_i \to Z_1$). The denominator measures how much of the variance in $X_1$ is explained by $Z_2, \ldots, Z_d$. For $k \in [2:d]$, we receive

$$\begin{aligned}\hat{\beta}_k &= \beta_k + \beta_1 \frac{\text{Cov}\left[Z_k, Z_1\right]}{\mathbb{V}\left[Z_k\right]} - \beta_1 \frac{1}{1-R^2} \frac{\text{Cov}\left[X_1, Z_k\right]}{\mathbb{V}\left[Z_k\right]} \left(\frac{\text{Cov}\left[X_1, Z_1\right]}{\mathbb{V}\left[X_1\right]} - \sum_{i=2}^{d} \frac{\text{Cov}\left[X_1, Z_i\right] \text{Cov}\left[Z_1, Z_i\right]}{\mathbb{V}\left[Z_i\right] \mathbb{V}\left[X_1\right]}\right) \\ &= \beta_k + \beta_1 \gamma_{Z_1,Z_k} - \beta_1 \frac{1}{1-R^2} \gamma_{X_1,Z_k} \left(\gamma_{Z_1,X_1} - \sum_{i=2}^{d} \gamma_{Z_i,X_1}\gamma_{Z_1,Z_i}\right) \\ &= \beta_k + \underbrace{\beta_1 \gamma_{Z_1,Z_k}}_{\text{omitted variable bias}} - \underbrace{\hat{\beta}_1 \gamma_{X_1,Z_k}.}_{\text{Correction since partially observed}}\end{aligned}$$

Instead of just encoding the effect of $Z_k$ on $Y$ and partial effect of $Z_1$ on $Y$ like before, the estimate $\hat{\beta}_k$ now also corrects for the fact that $Z_1$ is partially observed. The magnitude of the correction depends on the parameter estimate for the partially observed variable as well as the linear relationship between $X_1$ and $Z_k$.

## B.3 Supplementary Figures



(a) Female group

(b) Male group

Figure B.1: Test set $R^2$ over varying levels of feature missingness in the ACS Income data set. Results are reported as averages over 50 runs on the test set. Variation in results was minimal.

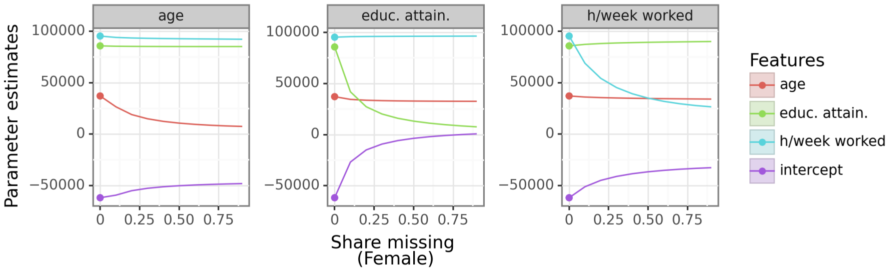


Figure B.2: Parameter estimates over varying levels of feature missingness in the ACS Income data set. Each panel indicates a different feature selected for missingness injection. Points indicate the true parameters from the semi-synthetic ground-truth model. Results are reported as averages over 50 runs. Variation in estimates over runs was minimal. Note that only estimates for continuous features are displayed and the estimated parameters for the levels of the categorical variables worker class, marital status, and relationship to reference person are omitted for readability.

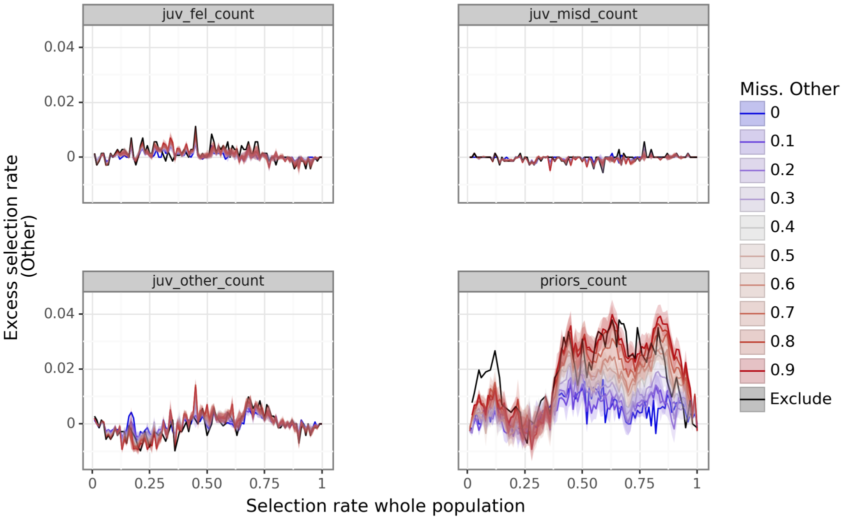


Figure B.3: Multiple imputation excess selection rate of racial group Other (i.e. not African-American) at different selection rates of the whole population with synthetic two-year recidivism outcomes using the COMPAS data set. Results are reported as averages over 30 runs on a test set. Shaded areas correspond to one standard deviation in each direction of the mean.

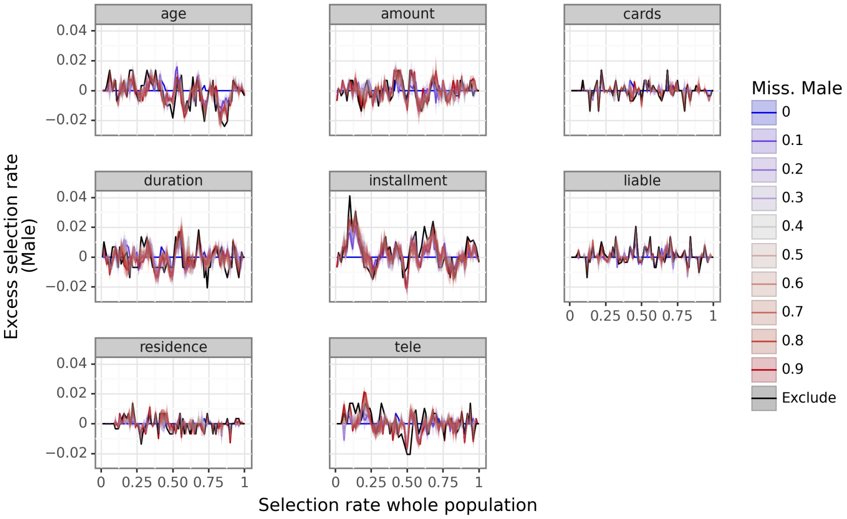


Figure B.4: Excess selection rate of male group at different selection rates of the whole population with synthetic outcomes using the German credit data set. Each panel represents a feature that has been corrupted by missingess in independent runs of the experiment. Feature missingness is added to the male group with 0-90% missing in 10 percentage point increments. The black curves show performance when excluding the whole feature column from modeling. Results are reported as averages over 50 runs on a test set. Shaded areas correspond to one standard deviation in each direction of the mean. Note that feature missingness is only injected into continuous features and models are estimated using the displayed features as well as the available categorical features checking account status, credit history, purpose, savings, employment, marital status, type of owned property, other installment plans, housing type, and job type.

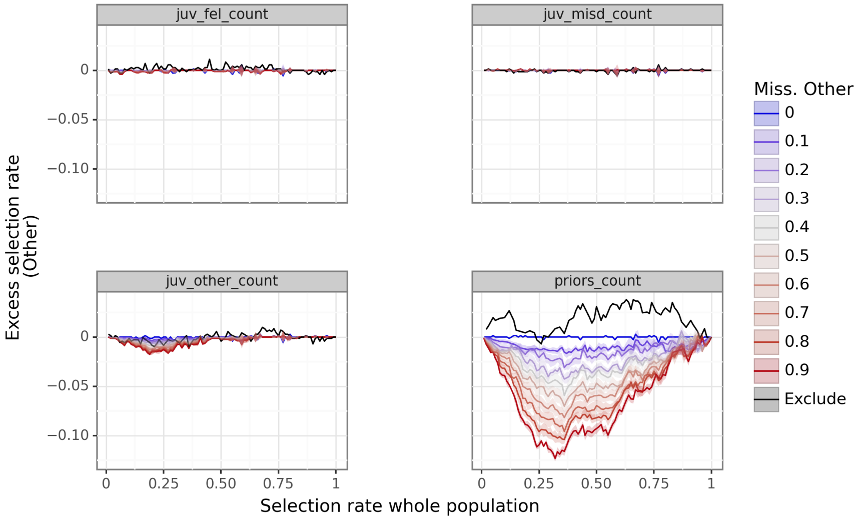


Figure B.5: Excess selection rate of racial group Other (i.e. not African-American) when training on rows without 0-entries at different selection rates of the whole population with synthetic two-year recidivism outcomes using the COMPAS data set. Results are reported as averages over 30 runs on a test set. Shaded areas correspond to one standard deviation in each direction of the mean.

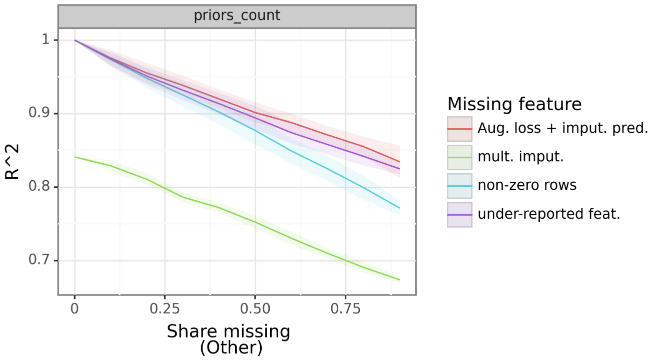


Figure B.6: Test set $R^2$ of different solution approaches using the COMPAS data set with synthetic two-year recidivism outcomes. Under-reporting is injected into the feature ‘priors count’ of group ‘Other’ (i.e. not African-American). Results are reported as averages over 30 runs. Shaded areas correspond to one standard deviation.

## B.4 Additional experiments and results

### B.4.1 Beyond the noise-free setting

**Motivation** The experiments on the publicly available data sets discussed in Sections 3.8 and 3.9 rely on semi-synthetic outcomes that are computed as deterministic linear functions of correctly measured features. The implicit simplifying assumptions are that, without feature missingness, a linear model on the data can retrieve the true data generating model $f(X) = \alpha + \beta^T Z$ and the exact outcomes $Y$ as recorded in the data. This modeling choice facilitates isolation of the effect of unobserved feature missingness by explicitly excluding potential effects of model misspecification and regression noise. In real-life applications, we can generally not predict outcomes exactly even if correctly measured features are available. In the following additional set of experiments, we loosen the assumption of a noise-free regression setting to allow for more general settings.

**Experimental setup** We follow a similar experimental setup as described in Section 3.8.2. Instead of relying on noise-free regression labels $Y = \alpha + \beta^T Z$, we add some noise back into the system by setting

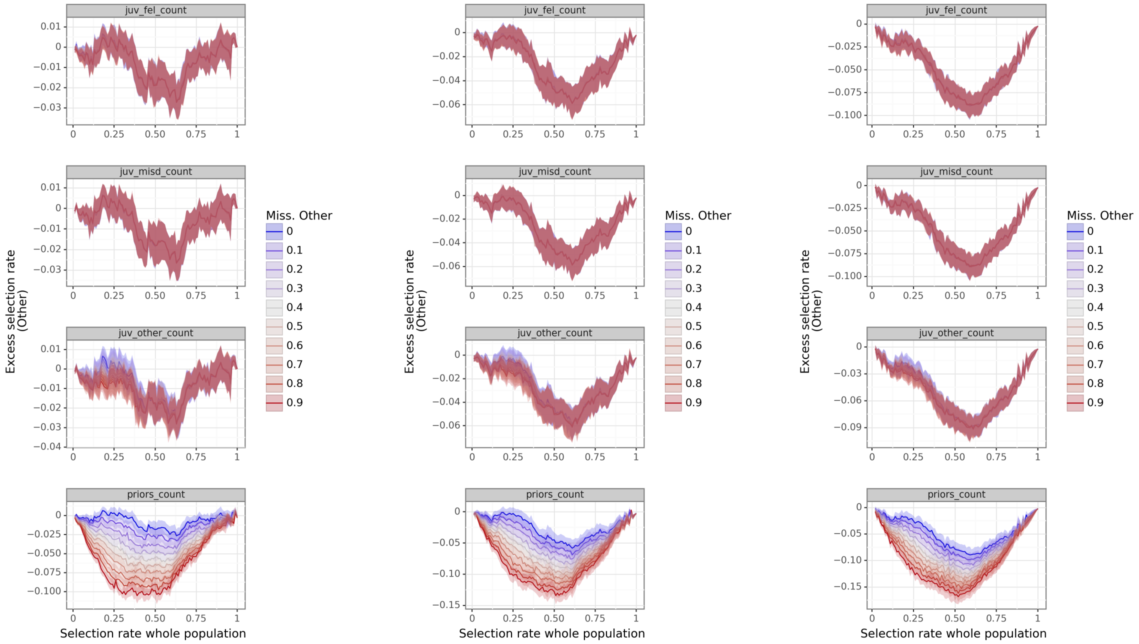
$$Y = \alpha + \beta^T Z + \varepsilon.$$

Here, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. and assumed to have mean zero. Like before, fitting a linear regression of $Y$ on features $Z$ with sufficient data yields the correct parameter estimates $\hat{\beta} = \beta$ and $\hat{\alpha} = \alpha$. However, in contrast to the noise-free setting, the prediction model $\hat{Y}_Z = \alpha + \beta^T Z$ can only retrieve labels $Y$ up to random noise. The $R^2$ of this prediction model can be controlled via the variance $\sigma^2$ by setting

$$\sigma^2 = \frac{1 - R^2}{R^2} \mathbb{E}\left[ \left( (\alpha + \beta^T Z) - \mathbb{E}\left[ \alpha + \beta^T Z \right] \right)^2 \right].$$

We experiment with $R^2$ values between 0.1 and 1.0 in increments of 0.1. Instead of comparing predictions $\hat{Y}_Z$ to predictions under unobserved feature missingness $\hat{Y}_X$, we compare thresholded versions of outcomes $Y$ and $\hat{Y}_X$ directly to measure both the impact of missingness and regression noise.

**Results** Figure B.7 depicts a subset of the results for the COMPAS data set. Comparing against the results of the noise-free setting summarized in Figure 3.3, we observe that the group Other is under-selected to a greater extend with additional noise. Under-selection occurs even if no feature missingness is added (dark blue curves) and increases with increasing noise, i.e. decreasing $R^2$ of the model on $Z$. On a high level, this occurs because the predictions $\hat{Y}_Z$ concentrate more closely around their group-level means as compared to the true values $Y$. The mean of $\hat{Y}_Z$ is smaller for the group Other than the group



(a) Model on $Z$: $R^2 = 0.9$.  (b) Model on $Z$: $R^2 = 0.6$.  (c) Model on $Z$: $R^2 = 0.3$.

Figure B.7: Excess selection rate under missingness over true labels $Y$ with different levels of $R^2$ for the model on correctly measured features $Z$. Low $R^2$ indicates a high level of noise and vice versa. Missingness injected into the features of group Other (i.e. not African-American) in the COMPAS data set. Results are reported as averages over 30 simulation runs with shaded areas representing one standard deviation in each direction.

African-American which leads to under-selection of the group Other as compared to the true $Y$ at many thresholds. We note that the group-level variances in outcomes $Y$ and predictions $\hat{Y}_Z$ play a role in this dynamic as well.

The isolated effect of unobserved feature missingness in the studied setting appears to be similar to the effect in the noise-free setting. As missingness is introduced into the group Other via the feature ‘priors count’, the group Other is further under-selected. The more missingness is injected, the more the group is under-selected. The magnitude of under-selection due to missingness is comparable across different levels of regression noise.

Overall, the results give us some insight into what to expect in more realistic settings of unobserved feature missingness. Instead of selection rate disparities that are exclusively due to differential feature under-reporting, disparities in the studied setting also depend on regression noise which, together, leads to increased disparities overall.

### B.4.2 Possibility of decreasing disparities

We conduct our main experiments on three publicly available data sets, i.e. COMPAS data (Angwin et al., 2016), German credit data (Repository, 1994), and ACS Income data (Ding et al., 2021), where each numerical feature is considered for the effect of unobserved missingness. As discussed in Section 3.9, the results suggest that, if an effect is present, unobserved feature missingness generally leads to under-selection of the group with missingness which aligns with Case 2 from the theoretical derivations in Section 3.6. If the group with missingness aligns with the group that is less frequently selected in the ground-truth model, this implies that differential feature under-reporting leads to increased selection rate disparities.

All three data sets have a numerical age feature which was considered for missingness but omitted for the discussion of results in the main text. In contrast to most other features (e.g. the counts in the COMPAS data), the default value of 0 is somewhat unintuitive for age and lies outside of the feature's support in each of the data sets. Studying the effect of fitting a model on differentially available data directly is less compelling in this setting since we essentially have indicators for missingness and could hope to use missing data methods like imputation directly. Nevertheless, we discuss the results for unobserved feature missingness in age for the COMPAS data set in the following as it presents the only empirical example for decreasing disparities we encounter in our experiments.

Figure B.8 depicts the parameter estimates and excess selection rate of group Other (i.e. not African-American) when fitting a model on data with unobserved feature missingness in the feature 'age' for group Other. We see that missingness in this setting leads to over-selection of the group with missingness. This over-selection is increasing with increasing levels of missingness. As the figure shows, the regression parameter for age is negative with an attenuation effect when missingness is injected. This means that in the semi-synthetic ground truth model and in the prediction models under missingness younger defendants are more likely to reoffend than older defendants. The feature correlations between age and juvenile crime counts (felony, misdemeanor, and other) are negative in the data while the correlation between age and the feature 'priors count' is positive. This leads to parameter estimates that are increasing for increasing missingness in age for juvenile crime counts and decreasing for increasing missingness for priors count exactly as predicted by the theoretical analyses in Proposition 6. Ultimately, this example shows how, in some settings, disparities may decrease as a function of unobserved missingness which aligns with

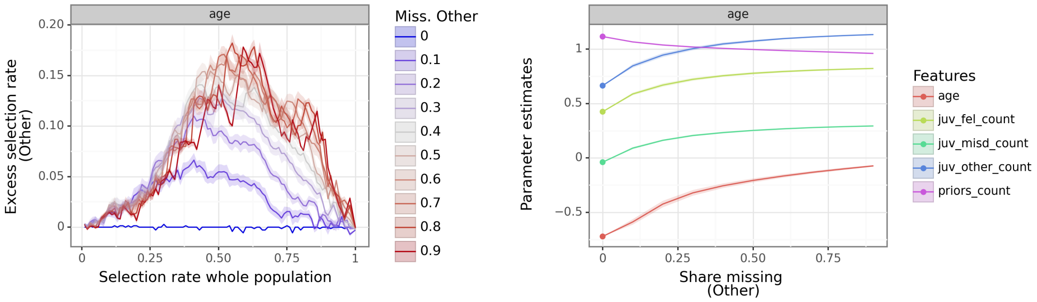


Figure B.8: Excess selection rate of group Other (i.e. not African-American) at different population selection rates with synthetic outcomes using the COMPAS data set, and the respective parameter estimates. Missingness is added to the feature ‘age’ in group Other. Results are reported as averages over 30 simulation runs with shaded areas representing one standard deviation in each direction. Note that parameter estimates are only displayed for continuous count features and age to preserve readability. The models additionally take sex and the categorical feature charge degree into consideration.

Case 1 from the theoretical discussions in Section 3.6. However, the example presented here is somewhat artificial and we find that typically disparities are increasing with differential feature under-reporting.

# Appendix C

# Appendix for Chapter 4

## C.1 Supplementary figures for in-processing Equalized Odds intervention

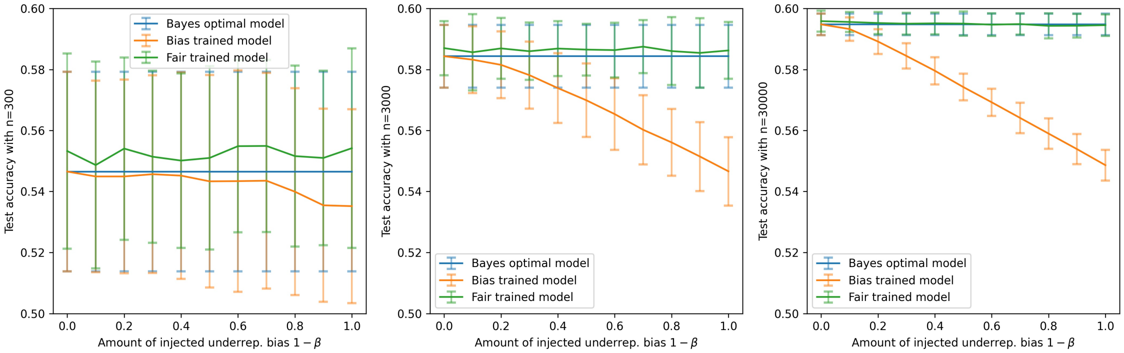


Figure C.1: Average test accuracy of the data-driven Bayes optimal model, model trained on biases data, and model trained on biased data with intervention over different amounts of injected under-representation bias. Results are averaged over 50 simulation runs and a total of $n \in \{300, 3000, 30000\}$ examples were used for training and testing each (left to right). Note that $\eta = 0.4$ and thus the data-driven Bayes optimal classifier reaches an accuracy of 0.6 if sufficient training data are used. We see that performance is only close to 0.6 if $n = 30000$ examples are used for training. We further see that the fair learned model outperforms the bias trained model for all $n \in \{300, 3000, 30000\}$.

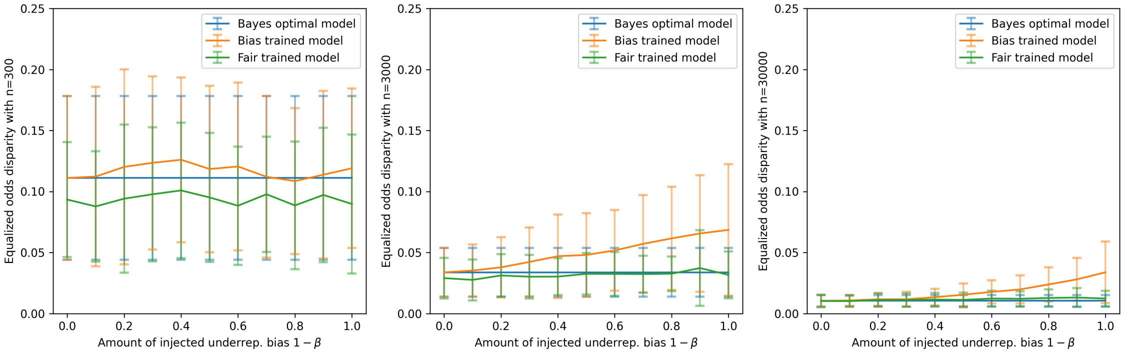


Figure C.2: Average test Equalized Odds disparity of the data-driven Bayes optimal model, model trained on biases data, and model trained on biased data with intervention over different amounts of injected under-representation bias. Results are averaged over 50 simulation runs and a total of $n \in \{300, 3000, 30000\}$ examples were used for training and testing each (left to right). We see that the fair learned model diminishes the Equalized Odds disparity for all $n \in \{300, 3000, 30000\}$. With sufficient data, the Bayes optimal model reaches a disparity of zero as expected in which case it aligns almost perfectly with the fair trained model ($n = 30000$).

## C.2 Results for post-processing Equalized Odds intervention

The results described in Section 4.3 and 4.4 are based on in-processing Equalized Odds fairness intervention following the reductions approach from Agarwal et al. (2018a). This is a natural choice as Blum and Stangl (2020) call for fairness constrained empirical risk minimization for their findings, one of which we are replicating on Section 4.3. However, since post-processing methods are desirable in some settings, we repeat the same experiments with the threshold based post-processing Equalized Odds algorithm from (Hardt et al., 2016). The results are presented and compared to the in-processing method in the following.

#### C.2.0.1 Section 4.3: Case study

Figure C.3 displays the test set fidelities in the original setting of Blum and Stangl (2020) for different amounts of training data $n \in \{300, 3000, 3000\}$ under post-processing Equalized Odds intervention. For consistency, we choose the same parameter inputs as in Section 4.3.3, i.e. $r = 0.2, \eta = 0.4$, and $R = 50$. The figure suggests that $n = 300$ and $n = 3000$ examples for training and fairness thresholding are not sufficient to retrieve the Bayes optimal classifier. In fact, Figure C.4 confirms that the data-driven Bayes optimal models are far from the analytical Bayes optimal model (which has accuracy of 60%) for $n \in \{300, 3000\}$. For $n = 30000$, the post-processing method successfully classifies like the Bayes optimal classifier in most cases and for most levels of bias $1-\beta$. We note that the post-processing method for Equalized Odds intervention requires a positive number of predictions for all classes in each of the groups in order to debias the predictions. For small data sets with a lot of injected bias, this assumption is not met in some cases which renders the post-processing method not applicable.

We compare the post-processing results from Figures C.3-C.5 to the analogous figures from the in-processing Equalized Odds method (see Figures 4.3,C.1 and C.2), and observe very similar results. This similarity is unsurprising given the general characteristics of the in-processing and post processing Equalized Odds solutions. On a high level, the threshold based post-processing method from Hardt et al. (2016) takes the predictions of a biased classifier and debiases them to fulfill Equalized Odds. This leads to an unbiased classifier which is not necessarily guaranteed to be the fair classifier with minimal error. In contrast, the reductions Equalized Odds approach from Agarwal et al. (2018a) realizes different trade-offs between fairness and accuracy from which the user can select and our implementation chooses to weigh the factors 50/50. In the given setting, Blum and Stangl (2020) show that there is no trade-off between fairness and accuracy and the classifier with minimum error on the test set aligns with the classifier with minimal Equalized Odds violation.

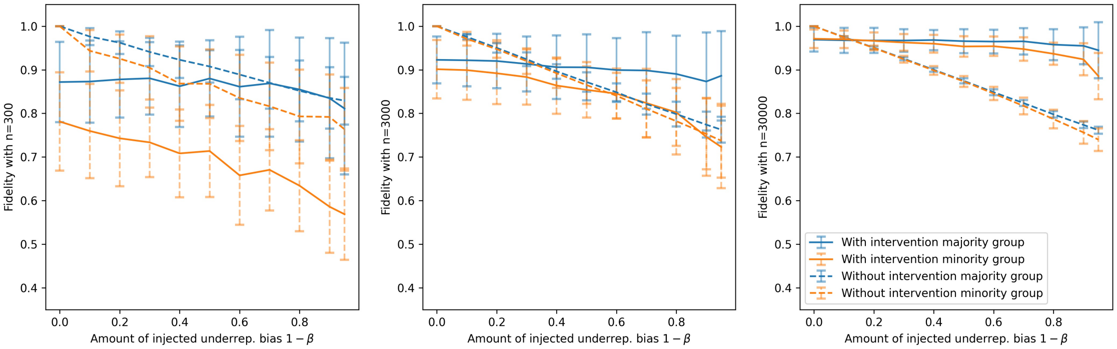


Figure C.3: Test set fidelity between Bayes optimal classifier and models trained on biased data with and without post-processing fairness intervention using $n = 300, 3000, 30000$ (left to right) samples for training and testing each. Experiments are repeated 50 times and reported as averages with one standard deviation in each direction. We note that the post-processing method requires a positive number of biased predictions for all outcomes in each of the groups which was violated in 5 (8) cases for 90% (95%) bias injection with $n = 300$. These cases are excluded from the figure.

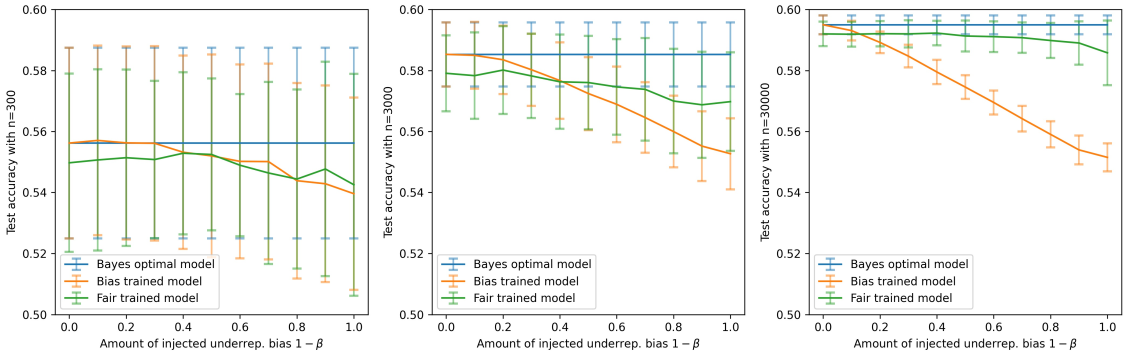


Figure C.4: Average test accuracy of the data-driven Bayes optimal model, model trained on biases data, and model trained on biased data with post-processing intervention using $n = 300, 3000, 30000$ (left to right) samples for training and testing each. Experiments are repeated 50 times and reported as averages with one standard deviation in each direction. We note that the post-processing method requires a positive number of biased predictions for all outcomes in each of the groups which was violated in 5 (8) cases for 90% (95%) bias injection with $n = 300$. These cases are excluded from the figure.

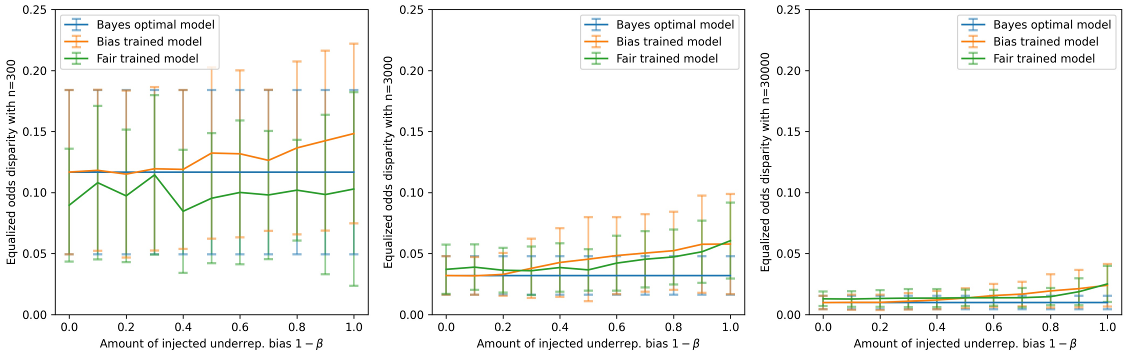


Figure C.5: Average test Equalized Odds disparity of the data-driven Bayes optimal model, model trained on biases data, and model trained on biased data with post-processing intervention using $n = 300, 3000, 30000$ (left to right) samples for training and testing each. Experiments are repeated 50 times and reported as averages with one standard deviation in each direction. We note that the post-processing method requires a positive number of biased predictions for all outcomes in each of the groups which was violated in 5 (8) cases for 90% (95%) bias injection with $n = 300$. These cases are excluded from the figure.

#### C.2.0.2 Section 4.4: Exploration

While in-processing and post-processing Equalized Odds intervention showed to lead to similar results in replication of Theorem 13, there are some differences in observations for our exploratory experiments. We discuss these differences in the following.

Figure C.6 depicts test set fidelity after post-processing intervention when $1 - \beta = 0.4$ under-representation bias is injected and base rates differ which corresponds to the setting from Section 4.4.1 and Figure 4.4. While the qualitative observation, i.e. the Bayes optimal classifiers can only be retrieved if base rates are the same across groups, is the same for both in-processing and post-processing intervention, the shapes of the curves differ when base rates diverge. Overall, the minority group fidelity appears to be lower in the post-processing case than in the in-processing case. We hypothesize that this occurs because, for non-zero base rate differences, the in-processing method trades off some amount of fairness for accuracy while the post-processing method does not have this option as discussed in Appendix C.2.0.1.

In the cases of sampling bias and label bias (Figure C.7 and C.8), we receive similar results for in-processing and post-processing interventions in experiments with the same base rates across groups. When base rates differ, the in-processing method outperforms the post-processing method in fidelity with the same reasoning as above.

We depict the post-processing results for feature measurement bias in Figure C.9 and observe considerable performance differences when comparing to the in-processing results from Figure 4.7. While the Bayes optimal classifiers are approximated closely for most bias levels when applying the in-processing Equalized Odds intervention, the post-processing intervention leads to decreasing fidelity in one of the groups as more bias is injected. To understand why this is the case, we recall our reasoning from Section 4.4.4. As more and more information is removed from the features, the predicted minority group scores of the models concentrate more around their means. This introduces a small amount of disparity into the Bayes optimal predictions on the biased training set which is accepted by the in-processing intervention in favor of higher accuracy. In contrast to the in-processing intervention, the post-processing method cannot trade fairness and accuracy off and instead rebalances predictions on the biased training data set to fulfill Equalized Odds exactly. This leads the model to select group-specific thresholds with poor performance on the unbiased test data which is reflected as decreases fidelity here.

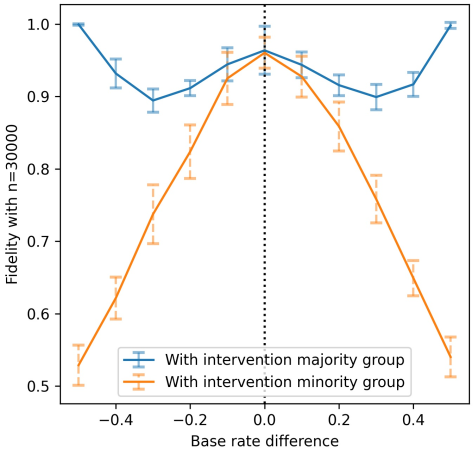


Figure C.6: Test set fidelity between Bayes optimal classifier and model trained on biased data with Equalized Odds post-processing intervention. Results are reported as an average over 50 simulation runs. Error bars correspond to one standard deviation in each direction.

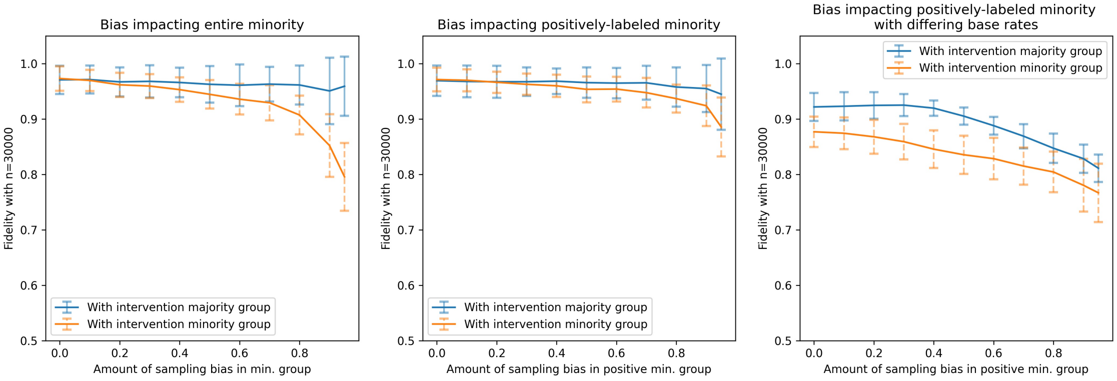


Figure C.7: **Sampling bias.** Test set fidelity between Bayes optimal classifier and models trained on biased data with Equalized Odds post-processing intervention on 30000 samples for training and testing each. Results are reported averaged over 50 simulation runs with error bars for one standard deviation in each direction. Bias is injected into either the entire minority group (left), or the positively labeled minority group (middle). On the right, bias is injected into the positively labeled minority group and we assume a base rate difference of -0.2

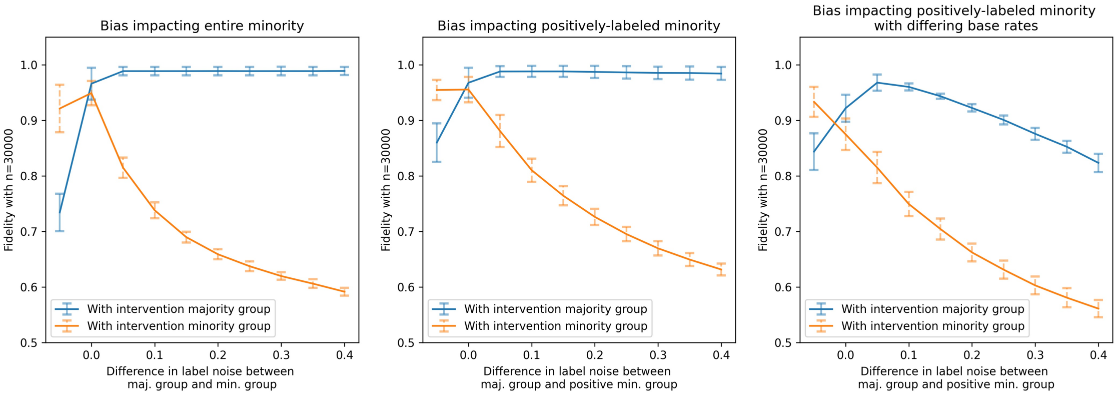


Figure C.8: **Label bias.** Test set fidelity between Bayes optimal classifier and models trained on biased data with Equalized Odds post-processing intervention on 30000 samples for training and testing each. Results are reported averaged over 50 simulation runs with error bars for one standard deviation in each direction. Bias is injected into either the entire minority group (left), or the positively labeled minority group (middle). On the right, bias is injected into the positively labeled minority group and we assume a base rate difference of -0.2.

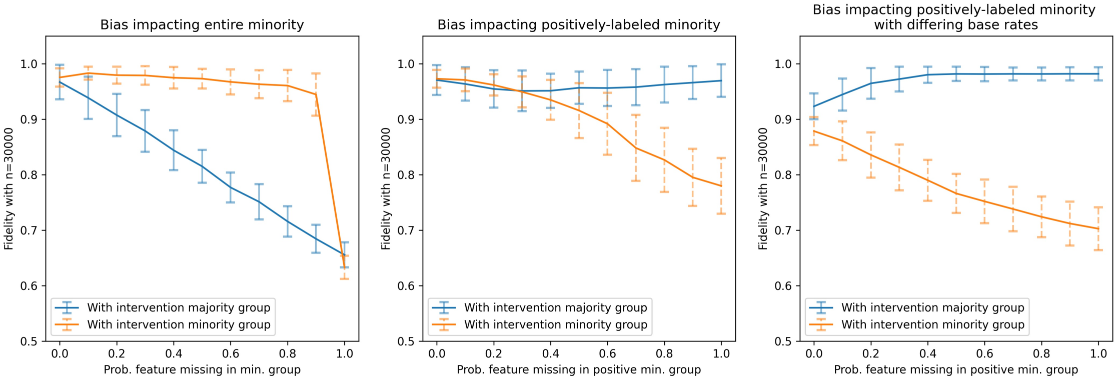


Figure C.9: **Feature measurement bias.** Test set fidelity between Bayes optimal classifier and models trained on biased data with Equalized Odds post-processing intervention on 30000 samples for training and testing each. Results are reported averaged over 50 simulation runs with error bars for one standard deviation in each direction. Bias is injected into either the entire minority group (left), or the positively labeled minority group (middle). On the right, bias is injected into the positively labeled minority group and we assume a base rate difference of -0.2.

## C.3 Experiments with confounding bias

**Motivation**. Real-world applications typically operate on more complex data sets than the synthetic data assumed in the main text of this study. We extend our explorations into the direction of more realistic data settings by conducting a sequence of experiments on data impacted by confounding bias. Furthermore,

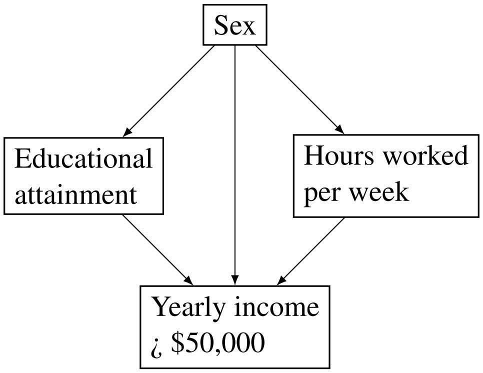


Figure C.10: Directed acyclic graph for simulated income data. The variable sex confounds the relationship between educational attainmenet and yearly income, as well as the relationship between hours worked per week and yearly income.

we use this additional set of experiments to illustrate how the proposed fairness sandbox can be used to compare interventions with different fairness metrics in a user-specified setting.

**Data set**. In the case of real-world applications, training data can rarely be assumed to be free of biases. While a thorough assessment of the applicability of the counterfactual bias injection framework in real-world settings is an important avenue for future work, we concentrate on synthetic data at this time. We assume a binary income prediction task, i.e. income above or below \$50,000 per year, and synthetically generate features, group membership information, and outcomes based on the causal graph depicted in Figure C.10. The causal graph and parameters for data generation are inspired by real 2018 US Census data from California obtained through the Folktables package (Ding et al., 2021). Synthetic data is generated using the MirrorDataGenerator package (Yang, 2023), showcasing the ease with which different software packages (e.g., for syntehtic data generation) can be integrated into the sandbox pipeline.

More concretely, the data generating mechanism from Figure C.10 is specified by the following components.

1. **Sex:** Sex is Bernoulli random variable with $P(\text{male}) = 0.53$ and female otherwise which is based on the group shares in the real Census data set.
2. **Educational attainment:** Educational attainment is an ordinal variable with values between $1$ and $24$ broadly referencing the number of years of schooling an individual has obtained. This also includes college, graduate school, etc. We sample educational attainment from a group-dependent Gaussian $\mathcal{N}(\mu_G, \sigma_G^2)$ where $G \in \{\text{male}, \text{female}\}$. Values outside of the interval $[1, 24]$ are mapped

to the end-points. In compliance with the real US census data, we set $\mu_{\text{male}} = 18.23$, $\mu_{\text{female}} = 18.74$, $\sigma_{\text{male}} = 4.11$, and $\sigma_{\text{female}} = 3.72$. Note that, on average, female individuals in the data have slightly higher education attainment.

3. **Hours worked per week:** Similar to educational attainment, the number of hours worked per week is sampled from a group dependent Gaussian $\mathcal{N}(\mu_G, \sigma_G^2)$ where $G \in \{\text{male}, \text{female}\}$. Here, the census data suggests $\mu_{\text{male}} = 40.05$, $\mu_{\text{female}} = 35.43$, $\sigma_{\text{male}} = 12.7$, and $\sigma_{\text{female}} = 12.94$. Note that, on average, female individuals work about 4.5h less per week than their male counterparts.

4. **Yearly income ¿$50,000:** We assume that the yearly income of an individual is a linear function of educational attainment, hours worked per week, and sex. More concretely, we set

$$\begin{aligned}\text{yearly income in \$ } =& 1100.43 \times \text{educ. attain} \\ &+ 682.84 \times \text{hours worked per week} \\ &+ 3805.54 \times 1\,(\text{sex} = \text{Male})\,.\end{aligned}$$

   The exact parameters in this toy model are retrieved from the real Census data by (1) fitting a logistic regression model to the data, (2) training a linear regression on the scaled probability predictions of the logistic regression, and (3) extracting the linear regression parameters. For the subsequent binary prediction task, we threshold the yearly income values at %50,000.

**Sandbox and experiment setup**. We consider the effects of confounding bias and fairness intervention in the synthetic income data setting by comparison across three types of models.

1. **Unconfounded model:** Assume we have full access to the data-generating process and fit a prediction model that accounts for the confounding variable sex. The features in this model are education attainment, hours worked per week, and sex.

2. **Confounded model:** Assume we do not have access to the confounding variable sex. The features in this model are educational attainment and hours worked per week.

3. **Model with fairness intervention:** Assume we are in the setting of the confounded model, but we are intervening on the fairness of the trained model. Fairness intervention is this setting is conducted using the Grid Search function from Fairlearn with different fairness metric specifications. The

methods here are based on the reductions approach from Agarwal et al. (2018a). We repeat the experiments for the following intervention types: (1) Equalized Odds, (2) Equality of Opportunity, (3) Demographic Parity.

We chose logistic regression as the function class for all prediction models. A total of 10,000 data points is synthetically generated using the described data generation mechanism. The data is subsequently separated in 50% training and 50% test data.

Similar to before, results are visualized and evaluated using various performance metrics. All evaluations are provided on a hold-out test set and averaged over several simulation runs where a new data set is sampled for every run.

**Results**. Experimental test set results are summarized in Figures C.11 and C.12. We first note that the unconfounded model reaches an almost optimal average test accuracy of 99.97% which is unsurprising given the synthetic data generation process. Since the fairness metrics Equalized Odds and Equality of Opportunity are conditional on true outcomes, there is almost no fairness violation with respect to these metrics in the unconfounded model (Fugure C.12). The Demographic parity constraint however is agnostic to differences in the true prevalence of incomes over $50,000 across groups which leads to an initial test set disparity of 13.38% on average. This disparity corresponds roughly to the difference in true rates of incomes over $50,000.

Figure C.11 shows that the confounded model has a median test set accuracy of 92.75%. Confounding leads to a slightly lower accuracy for the minority group as compared to the majority group, as well as increased fairness violation in terms of Equalized Odds and Equality of Opportunity. On a high level, the confounded model can only access the information of the sex variable through the features education attainment and hours worked per week which are deferentially predictive for the two groups. In this example, confounding leads to over-prediction of high incomes for the female group (false positives) and under-prediction of high incomes for the male group (false negatives) which implies violations of Equalized Odds and Equality of Opportunity. At the same time, these error types lead to a more even distribution of rates of predicted incomes over $50,000 across groups which implies a decreased Demographic Parity error (Figure C.12.

Fairness intervention in the confounded model considerably influences both accuracy and fairness constrain violation metrics. Both Equalized Odds and Demographic Parity intervention lead to significantly decreased test set accuracy with 80.16% median accuracy for Equalized Odds and 89.13% for Demographic Parity. The median test set accuracy with Equality of Opportunity intervention reaches 92.71% which is comparable with the accuracy of the confounded model. Compared to the confounded model, the Equality of Opportunity model decreases test set fairness disparities in terms on both Equalized Odds and Equality of Opportunity. We note that the in-processing reductions approach to enforcing fairness metrics in the experiments allows the models to trade-odd fairness and accuracy.

Both Equalized Odds and Equality of Opportunity interventions lead to increased Demographic parity violations which is in line with general impossibility results from the algorithmic fairness literature (Chouldechova, 2017; Kleinberg et al., 2017). In fact, in the studied setting increased Demographic Parity violation could be seen as indication of moving closer to the unconfounded model which has the highest violation among all the considered models.

Overall, our synthetic income data experiments show that confounding can introduce considerable fairness concerns in terms of Equalized Odds and Equality of Opportunity. In-processing fairness intervention with Equality of Opportunity showed promising improvements in terms of fairness without impacting prediction accuracy too much.

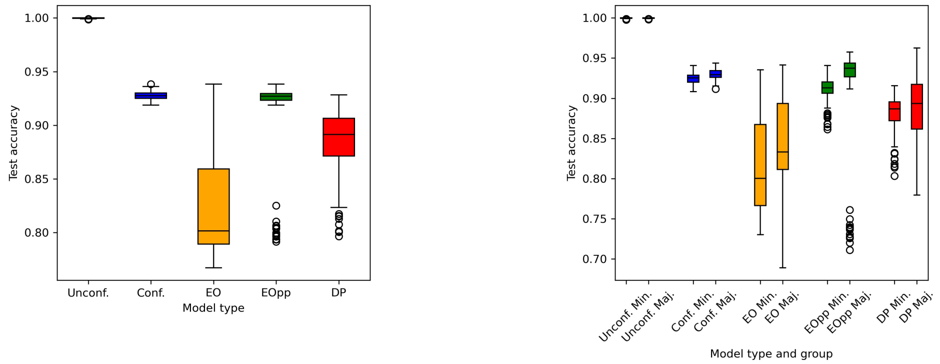
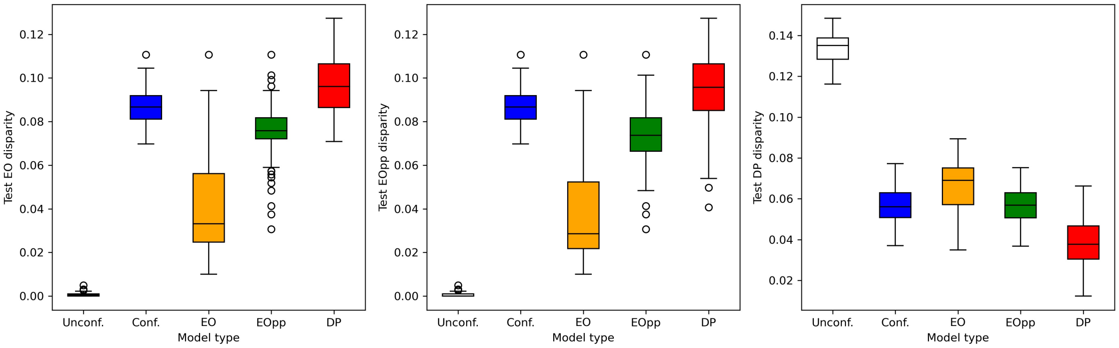


Figure C.11: Test set accuracy of models without confounding, with confounding, and with confounding and fairness intervention (EO = Equalized Odds, EOpp = Equality of Opportunity, DP = Demographic Parity). The left plot shows the overall accuracy while the right plot displays accuracy by group (Maj. = Male, Min. = Female). Results are depicted for 100 simulation runs.



(a) Equalized Odds

(b) Equality of Opportunity

(c) Demographic Parity

Figure C.12: Test set fairness violation of models without confounding, with confounding, and with confounding and fairness intervention (EO = Equalized Odds, EOpp = Equality of Opportunity, DP = Demographic Parity). Fairness disparity is measured in terms of Equalized Odds (left), Equality of Opportunity (middle), and Demographic Parity (right). Results are depicted for 100 simulation runs.

# Appendix D

# Appendix for Chapter 5

## D.1 Supplementary figures

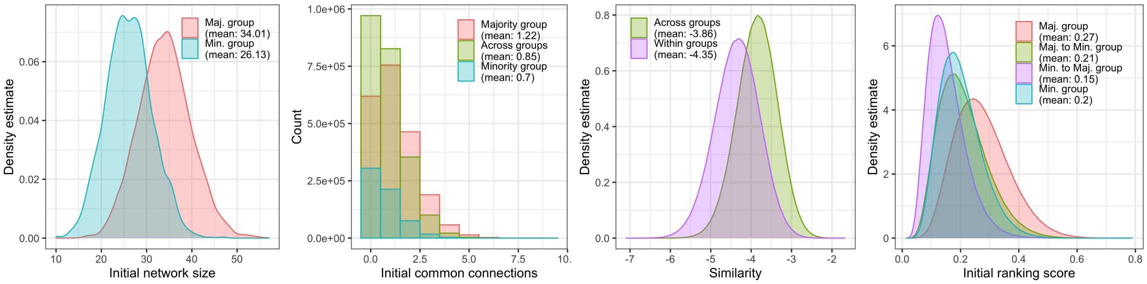


Figure D.1: Features of initial networks and ranking scores at $\mathbf{t = 0}$ for 10 simulation runs. Networks are initialized with a stochastic block model with $\mathbf{p}_{\text{connect}}(\mathbf{i}, \mathbf{j}) = \mathbf{0.04}$ for $\mathbf{i}, \mathbf{j} \in \mathbf{G_0}$ (majority group), $\mathbf{p}_{\text{connect}}(\mathbf{i}, \mathbf{j}) = \mathbf{0.032}$ for $\mathbf{i}, \mathbf{j} \in \mathbf{G_1}$ (minority group) and $\mathbf{p}_{\text{connect}}(\mathbf{i}, \mathbf{j}) = \mathbf{0.023}$ otherwise. We see that the initial network sizes tend to be larger for majority group members (top left), and members in the same groups tend to be more similar than members in different groups (bottom left). The average number of initial common connections between members in $\mathbf{G_0}$ ($\mathbf{G_1}$) is 1.22 (0.7) while the average number for member pairs across groups is 0.85 (top right). We use the scoring model to compute initial ranking scores between all unconnected members in $\mathbf{t = 0}$ and see that scores tend to be higher in the majority group as compared to the minority group and for in-group pairings of members as compared to pairings across groups (bottom right).

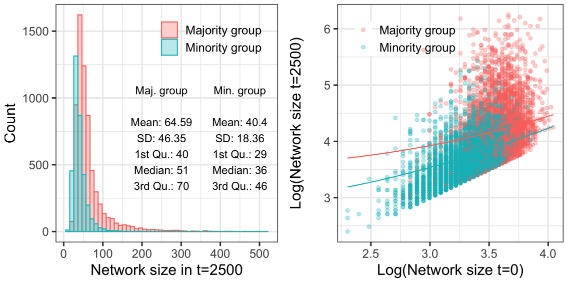


Figure D.2: Distribution of network sizes in $\mathbf{t} = \mathbf{2500}$ without fairness intervention (left). Log-log plot of network sizes without fairness intervention in $\mathbf{t} = \mathbf{0}$ and $\mathbf{t} = \mathbf{2500}$ with curves denoting the counterfactual network size at $\mathbf{t} = \mathbf{2500}$ if the total increase within groups was distributed evenly (right). Results include 10 simulation runs. We see that unconstrained recommendation leads to a group-wise rich-get-richer effect that benefits the majority group.

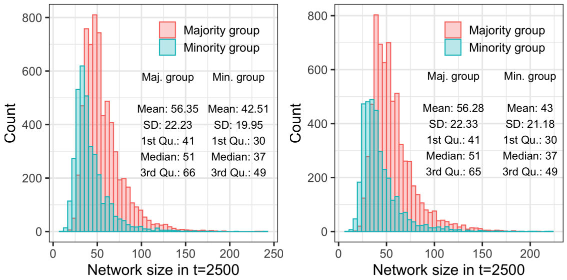


Figure D.3: Distribution of network sizes in $\mathbf{t} = \mathbf{2500}$ after demographic parity of exposure intervention (left) and after dynamic parity of utility intervention (right). Results include 10 simulation runs. We see that both interventions are unsuccessful in aligning network size distributions across groups.

## D.2 Comparison of main simulation and mixed preferential attachment model

The described Mixed Preferential Attachment (MPA) model is used to simulate a simplified mechanism of connection recommendation that qualitatively maintains many of the key aspects of the bias mechanism in our main simulation. First, we note that, although the main simulation assumes a fixed graph and new nodes are added at each iteration in the MPA model, recommendations are on average sought out more frequently by the members of the majority group in both models. The simulation study models this by exponential waiting times depending on the current network size of members, while the MPA model assumes a rate $r \leq 0.5$ of minority group members. Second, once a source member is selected, both models lead to connections to destination members based on two main features with similar interpretations. The scoring and connection models in the main simulation make use of the similarity between group-dependent features which renders members from the same groups more likely to be suggested and connect. Meanwhile, the MPA model uses the parameters in the mixing matrix to express a similar in-group preference when $p_0, p_1 > 0.5$. In addition, ranking scores in the main simulation are positively impacted by the number of common connections between the ranked members. This is not possible in the MPA model because source members enter the network without previous connections. However, the MPA model gives preference to destination members with large networks which goes into the same direction as the number of common connections and can lead to a similar effect (Liben-Nowell and Kleinberg, 2007). Overall, the models are similar enough that it is reasonable to expect that some of the qualitative observations we can make by analyzing the MPA model can be translated to insights into the behavior of the realistic simulation study on a group-aggregate level.

## D.3 Proof of Theorem 14

Our proof follows a similar procedure to the proof of Theorem 1 in (Avin et al., 2020), yet in our setting we are able to obtain a relatively simple closed form solution of the limit $\alpha$.

Note that exactly one new member and one connection are added in every time step. Given that the incoming source member is of group $G_i$, we denote the probability that the connection forms to a member of group $G_j$ by $P_{ij}$. We note that it holds $P_{ij} = 1 - P_{ii}$ for $i, j \in \{0, 1\}$ and use the mixing matrix $\pi$ compute

$$\begin{aligned} P_{00} &= (1-r)p_0 + (1-r)(1-p_0)P_{00} + rp_0P_{00} \\ \Leftrightarrow P_{00} &= \frac{(1-r)p_0}{r + p_0 - 2rp_0}, \end{aligned}$$

and

$$\begin{aligned} P_{11} &= rp_1 + r(1-p_1)P_{11} + (1-r)p_1P_{11} \\ \Leftrightarrow P_{11} &= \frac{rp_1}{1 - r - p_1 + 2rp_1}. \end{aligned}$$

Let $N_{t+1}$ be the number of group $G_1$ degrees added in step $t+1$. Then, it holds that

$$\begin{aligned}\mathbb{E}[N_{t+1}] &= 2rP_{11} + rP_{10} + (1-r)P_{01} \\ &= rP_{11} - (1-r)P_{00} + 1 \\ &= r\frac{rp_1}{1-r-p_1+2rp_1} - (1-r)\frac{(1-r)p_0}{r+p_0-2rp_0} + 1.\end{aligned} \tag{D.1}$$

We know that $\alpha_t = d_t(G_1)/d_t$ and $d_t = d_0 + 2t$ for all $t$. Thus,

$$\begin{aligned}\mathbb{E}[N_{t+1}] &= \mathbb{E}[d_{t+1}(G_1) - d_t(G_1)|\alpha_t] \\ &= \mathbb{E}[\alpha_{t+1}|\alpha_t]d_{t+1} - \alpha_t d_t \\ &= \mathbb{E}[\alpha_{t+1}|\alpha_t](d_0 + 2(t+1)) - \alpha_t(d_0 + 2t),\end{aligned}$$

and with we receive

$$\mathbb{E}[\alpha_{t+1}|\alpha_t] = \frac{\alpha_t(d_0+2t) + \mathbb{E}[N_t]}{d_0 + 2(t+1)} = \alpha_t + \frac{\mathbb{E}[N_t] - 2\alpha_t}{d_0+2(t+1)}.$$

Recursively inserting the conditional expected values of $\alpha_i$ for $i \in [t]$ and shifting $t$ by one gives

$$\begin{aligned}\mathbb{E}[\alpha_t] &= \alpha_0 \prod_{j=1}^{t}\left(1 - \frac{2}{d_0+2j}\right) + \sum_{i=1}^{t}\left(\frac{\mathbb{E}[N_t]}{d_0+2i}\prod_{k=i+1}^{t}\left(1-\frac{2}{d_0+2k}\right)\right) \\ &= \alpha_0 \prod_{j=1}^{t}\left(1 - \frac{2}{d_0+2j}\right) + t\frac{\mathbb{E}[N_t]}{d_0+2t}.\end{aligned}$$

Note that

$$\lim_{t\to\infty}\prod_{j=1}^{t}\left(1-\frac{2}{d_0+2j}\right) = 0$$

and thus with Equation (D.1)

$$\begin{aligned}\lim_{t\to\infty}\mathbb{E}[\alpha_t] &= \lim_{t\to\infty}\frac{\mathbb{E}[N_t]}{d_0/t+2} \\ &= \frac{\mathbb{E}[N_t]}{2} \\ &= \frac{1}{2}\left(r\frac{rp_1}{1-r-p_1+2rp_1} - (1-r)\frac{(1-r)p_0}{r+p_0-2rp_0} + 1\right).\end{aligned}$$

## D.4 Proof of Theorem 15

We use the same notation as in Appendix D.3 and note that in order for the limiting share $\alpha$ to be independent of $p_0$ and $p_1$, the same must be true for the conditional probabilities $P_{ij}$. Instead, the $P_{ij}$ must be chosen such that the expected number of $G_1$ balls added in $t$ fulfills

$$\mathbb{E}[N_t] = rP_{11} - (1-r)P_{00} + 1 = 2r,$$

since $\lim_{t\to\infty} \mathbb{E}[\alpha_t] = \mathbb{E}[N_t]/2$ (see proof of Theorem 14). This is trivially fulfilled by assuming $P_{00} = P_{11} = 1$ as described in Section 5.5.4. A more interesting solution is obtained by crafting an intervention which ensures that $P_{11} = r$ and $P_{00} = 1-r$ which we will do in the following.

We recompute the probabilities $P_{ij}$ as functions of the mixing matrix $\pi$ and the rejection sampling probabilities $q_{ij}$. In iteration step $t+1$, we have

$$\begin{aligned} P_{00} = {} & (\alpha_t - 1)q_{00}p_0 + (\alpha_t - 1)q_{00}(1-p_0)P_{00} \\ & + (\alpha_t - 1)(1-q_{00})P_{00} + \alpha_t q_{01}(1-p_0)P_{00} + \alpha_t(1-q_{01})P_{00}. \end{aligned}$$

Assume that $q_{ij} = 1 - q_{ii}$ for $i \neq j$. Setting $P_{00} = 1-r$ and solving for $q_{00}$, yields that we need to set

$$q_{00} = \frac{(1-r)(\alpha_t(p_0-2)+2)}{p_0(\alpha_t - r)}$$

as long as $\alpha_t \neq r$. Note that $p_0, p_1$ and $\alpha_t$ are bounded away from 0 and 1. Similarly, it holds that

$$\begin{aligned} P_{11} = {} & \alpha_t q_{11} p_1 + \alpha_t q_{11}(1-p_1)P_{11} + \alpha_t(1-q_{11})P_{11} \\ & + (1-\alpha_t)q_{10}p_1P_{11} + (1-\alpha_t)(1-q_{10})P_{11}. \end{aligned}$$

We set $P_{11} = r$ and receive

$$q_{11} = \frac{(1-\alpha_t)(1-p_1)r}{\alpha_t(p_1 - r) - p_1 r + r}$$

for $\alpha_t(p_1 - r) - p_1 r + r \neq 0$ and the claim follows.