Machine-Learning Fairness in Data Markets: Challenges and Opportunities

Roland Maio

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

Abstract

Machine-Learning Fairness in Data Markets: Challenges and Opportunities

Roland Maio

Machine learning promises to unlock troves of economic value. As advanced machine-learning techniques proliferate, they raise acute fairness concerns. These concerns must be addressed in order for the economic surpluses and externalities generated by machine learning to benefit society equitably. In this thesis, we focus on the economic context of data markets and theoretically study the impacts of intervening to achieve machine-learning fairness. We find that to effectively and efficiently intervene requires taking the data market into account in the design and application of the fairness intervention, i.e., how the intervention impacts the data market, how the data market impacts the intervention, and how their impacts interact. We study this interaction in two data-market settings to understand what information is necessary. We find that without taking into account the incentive structure and economics of a data market, fairness interventions can induce greater losses to efficiency than are necessary to achieve fairness—even potentially inducing market collapse. Yet, we also find that these losses can be recovered or even amortized away by suitably designing the intervention with appropriate information or under favorable market conditions. Overall, this thesis elucidates how data markets present both novel challenges and opportunities for machine-learning fairness. It demonstrates that efficiently intervening for machine-learning fairness can be more complicated in data markets—even infeasible! Excitingly, however, it also demonstrates that under favorable market conditions, fairness can be achieved at lower relative cost to efficiency than has previously been understood to

be possible. We hope that these initial theoretical findings ultimately contribute to the development of efficient and practical fairness interventions suitable for real-world application.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am eternally grateful to all the people who have supported me in any way on my PhD journey, big and small. The full circle is too large to enumerate here. Just figure that, if you are a being, I'm grateful to you.

Still, I would like to acknowledge a number of people explicitly here. First and foremost, my advisor Augustin Chaintreau. Thank you for believing in me, even when I didn't believe in myself. My labmates: Ana-Andreea Stoica, Yiguang Zhang, Xi Chen, and Gabriel Chuang, thank you for your wonderful company these past many years. My many co-authors and collaborators: Ryan Amos, Andrew Chong, Sean Cleary, Evan Dong, Jessie Finocchiaro, Thomas Krendl Gilbert, Sarah Hladikova, Prateek Mittal, Faidra Monachou, Carlos Mougan, Gourab K Patro, Manish Raghavan, Shubham Singh, Savannah Thais, Stratis Tsirtsis, Ayse Gizem Yasar, Juba Ziani and Miri Zilka, thank you all for inspiring me. And the entire ESP team, past present and future. In particular, Adam Cannon and Tim Randolph.

I wish for all of you only the best.

Dedication

To my wife Laísa, eu te amo.

Chapter 1: Introduction

Machine learning promises to unlock troves of economic value. One recent report estimates that just generative AI could potentially add $5.2 to $8.8 trillion annually [1][1]. The extent to which this potential is realized will depend on many factors including machine factors from hardware to software [2], human factors from the anatomy of work to worker training [1], and information factors from data to models. One of the most important and fundamental factors is data, specifically, the quantity and quality of data that are produced and used for machine learning.

Data is a fundamental determinant of machine-learning performance that can drive other determinants [3, 4]. Low quality data can adversely impact machine-learning performance [5]. Just increasing the amount of data can achieve new levels of state-of-the-art performance [6]. The performance gain often scales predictably in the amount of data; a practitioner can estimate the amount of data required to achieve a given level of performance [7]. And the capacity of models to absorb data often increases by increasing model complexity [8, 9]. Thus, data provide a lever that predictably and reliably improves performance—all by itself. This alone could drive strong demand for data, even before considering the impact of the fledgling generative-AI industry on data demand.

Generative-AI (genAI) technology companies consider data to be more than a production factor—data is a key source of competitive advantage [10, 11, 12]. In these early days, they are racing to acquire as much data as possible and to train models on ever larger amounts of data [10, 13, 14]. Their demand is so great that even all of the publicly-available and usable content on the internet may not be enough [10]; genAI companies may exhaust this supply by 2028 [15]. To increase the pool of data, genAI companies are also undertaking ethically-fraught activities such

[1]For perspective, the World Bank estimates that the United States' gross domestic product was $25.44 trillion in 2022.

as scraping copyrighted content without permission and updating their privacy policies to tap more of their users' data [13, 16]. Yet these activities can only increase the pool of data by so much and have their own challenges. Publishers and content creators are pushing back by filing lawsuits, blocking web crawlers, and negotiating licensing agreements for their content [17, 18, 19]. The limits and challenges of the existing stock of human-generated data may propel genAI companies to turn increasingly to data markets to satisfy their demand for data.

Data markets may provide an economic mechanism capable of meeting the surging demand for data. Data markets can be organized into many different market structures; they can be organized bilaterally, but also multilaterally with a coordinating central agent [19, 20, 21, 22]. Because raw data can be processed, data markets can trade in many kinds of data products [23, 24, 25, 26]. Among other possibilities, a data product can be as simple as a dataset [22]. But, they can also be derived in various ways from an underlying dataset. A derived data product can be the result of a SQL query [24], a notification of a specified change to the underlying dataset during a set time period [25], or predictions [26]. This ability to refine raw data has enabled a history of innovation to meet new needs and overcome transaction-hurdles. For example, privacy-preserving technologies enable data markets to unlock economic value and meet the privacy needs of various market participants including regulators, consumers, data sellers, and data buyers [27, 28]. Overall, data markets provide a flexible and evolving economic mechanism for coordinating the production and distribution of data; in tandem with unprecedented demand for data, this suggests that data markets may play a significant role in unlocking the potential economic value promised by machine learning. Yet, doing so will involve more than coordinating data production and applying machine learning—it will be inextricably intertwined with ethical concerns including legal and privacy, as we have seen, but also fairness.

It is now understood that machine-learning (ML) systems can raise fairness problems. Real-world fairness problems have been identified across a wide range of applications and domains including recidivism prediction in criminal justice [29], premium setting in auto insurance [30], online advertising for housing [31], semi-supervised learning in natural language processing [32],

and health-need prediction in health insurance [33]. These real-world examples illustrate a number of lessons. They highlight how fairness problems can arise from the complex interaction between stakeholders' interests, particularly in high-stakes applications. They suggest that a practitioner should expect fairness problems to arise in machine-learning applications as a matter of course [34]. And they indicate that these problems require technical solutions in addition to policy solutions. To meet the need to address fairness problems in machine learning, a vibrant research community has emerged.

The fair machine-learning community has made many significant contributions, in this thesis, we focus on two foundational contributions as critically related to addressing fairness in data markets. The first contribution is a framework for developing technical solutions based on three foundational scientific findings: 1) fairness cannot be ignored; unless fairness is explicitly addressed and controlled, ML systems will tend to be unfair in general [35]; 2) fairness can be operationalized in multiple ways that can come into irreconcilable conflict with each other [36]; and 3) building fair machine-learning systems can come into conflict with efficiency [37]. These findings provide the following guidelines for practitioners developing technical solutions, they must: 1) pro-actively assess fairness; 2) define fairness in a context-appropriate way that explicitly acknowledges and accepts inter-fairness tradeoffs; and 3) seek to achieve fairness as efficiently as possible. Guided by this framework, the community's second contribution is a wide array of *pure machine-learning solutions*.

Pure machine-learning solutions are an important contribution because they provide practitioners with tools that they can apply across a wide range of environments, including potentially data markets. Their distinctive feature is that they are domain agnostic, i.e., they abstract away domain-specific idiosyncracies that do not transfer seamlessly between domains and are assumed to be irrelevant to the machine-learning problem. Pure machine-learning solutions restrict their interventions and analyses to the standard framework of machine learning [38]; they focus on the data processing, learning algorithms, models, predictive performance metrics, and the machine-learning pipeline overall [39]. Yet the strength of pure machine-learning solutions is precisely

their weakness—extirpating a domain-specific idiosyncracy in a mathematical model does not extirpate the idiosyncracy in reality. That idiosyncracy may still affect the use of a pure machine-learning solution in ways that are unanticipated, adverse, and poorly understood. Data markets have a number of idiosynracies—such as scale and growth, incentives and strategic behavior, and economically-constrained decision-making—that may affect pure machine-learning solutions.

Altogether, the foregoing considerations raise intriguing questions about machine-learning fairness in data markets as data markets come to play an increasingly important role in unlocking the economic potential of machine learning. How will fairness problems arise in connection with data markets? When does it make sense to intervene in data markets as opposed to downstream machine-learning applications? Do pure machine-learning solutions applied in data markets behave as expected? Do the idiosynracies of data markets affect our understanding of fairness and pure machine-learning solutions? If not, how is their performance inconsistent with the literature and what new challenges do they face? If there are novel challenges, can they be overcome and what data-market specific information may be required to do so? On the flip side, do data-market aware interventions present new opportunities?

In this thesis, we theoretically explore these questions in two data market settings to provide some initial answers. In the first setting, we interpret the application environment of a pure machine-learning solution as a data market. This assumes ideal operating conditions and allows us to investigate the out-of-the-box application of that solution. In the second data-market setting, we revisit a well-known model of a data market and formulate a variant that captures fairenss and endogenizes data production and seller market participation. This allows us to study the economics of a fairness intervention and gain insight into the relationship between fairness and market conditions. In both of these settings, this thesis aims to establish two major findings.

First, we aim to demonstrate the need to design data-market aware fairness interventions. Without taking into account important aspects of data markets, fairness interventions may be inefficient. For example, data markets often seek to monetize data across multiple buyers, i.e. they seek to scale profits. Therefore, it is important for pure machine-learning solutions to scale profitably. In

Chapter 2 we examine an important pure machine-learning solution that is a prominent member of a class of solutions based on fairness-through-information-control, *fair representations* [40].[2] We show that fair representations are severely limited in their ability to scale profit; naively applying fair representations in a data market can incur a severe inefficency that can scale as a seller seeks to scale its profits. And importantly, inefficiency is not the only risk of pure machine-learning solutions—we show that failing to take the data market into account can present an even more severe risk.

Another important aspect of data markets is participation. Fundamentally, data markets are economic mechanisms for efficiently coordinating data production and distribution via the price mechanism [41]. Free markets form when participants are better off participating in the market compared to abstaining from the market. Although fairness interventions may be a moral imperative and intended to benefit some individuals or groups, they tend to increase transaction costs and lower total surplus. Standard analyses of pure machine-learning solutions do not take into account the underlying economics of the data market, in particular the impact of the intervention on market participation. We show in Chapter 3 that this can severely understate the impact of a fariness intervention: it is possible that a data market cannot bear a fairness intervention. Specifically, we find market conditions under which a data market forms in the absence of a fairness intervention but does not form when an intervention is undertaken because the cost of the intervention renders market participation irrational. Because no data is produced and no transactions occur when this happens, the fairness intervention can harm the very groups it was intended to benefit. Thus, fairness interventions in data markets may backfire. Notably, we find that this *backfire risk* cannot be controlled by modulating the intervention alone but depends on both the intervention and market conditions; every intervention backfires under certain market conditions and does not under others. Taken together, the inefficiency risk and the backfire risk indicate that achieving machine-learning fairness in data markets faces novel challenges.

Happily, the news is not all bad. The second major finding this thesis aims to demonstrate is that

[2] See Chapter 2 for details of fair representations and this class.

data markets also present opportunities for fairness interventions. We follow up on our result from Chapter 2 to demonstrate that that the inability to scale efficiently is not a fundamental limitation of the class of fairness-through-information-control solutions nor, indeed, of fair representations. We show that the inefficiency can often be fully recovered by formulating an economically-based variant of fair representations that we call *incentive-compatible fair representations*. In contrast to traditional fair representations, incentive-compatible fair representations model market participants as rational utility-maximizers that are responding to their incentives and take those incentives into account. Incentive-compatible fair representations turn out to be very flexible when they have complete information about all the market participants' incentives; this flexibility creates the opportunity to recover the inefficiency. Therefore, the challenges presented by data markets can be surmountable and, better yet, off-the-shelf pure machine-learning solutions may provide tractable starting points for designing efficient data-market aware fairness interventions.

Incentive-compatible fair representations took advantage of an opportunity presented by two important properties of data markets: 1) agent behavior is often rational; and 2) incentives shape rational actions. In Chapter 3 we show that additional important properties of data markets may present further opportunities for machine-learning fairness. We show that taking into account the endogeneity of supply—the fact that the amount of data that is brought to market depends on the market's ability to profitably extract economic value from the produced data—and market growth can change the dynamics of the cost of fairness in a data market. Specifically, the cost of fairness as a fraction of an agent's non-intervention utility can amortize towards zero as the market grows. This suggests that intervening for fairness may not be as costly in data markets under auspicious market conditions.

Overall, the goal of this thesis is to show that data markets merit further scrutiny by the fair machine-learning community. On the one hand, data markets present challenges that may render important pure-ML solutions unsuitable or non-performant in specific cases. On the other hand, these challenges may be surmountable, practitioners may be able to overcome these challenges with suitably-designed tools. Moreover, data markets present new opportunities for machine-

learning fairness. Our exciting results on the amortization of fairness suggest that achieving fairness in data markets may not be as costly as previously believed. This could significantly address one of the most important challenges to fairness interventions. And our work is only a start. There is much to do to build on these exciting and intriguing theoretical results towards supporting practical fairness interventions. And besides building directly on top of the results in this thesis, there are also many more important properties of data markets to investigate.

At the same time, our excitement should also be tempered. Our theoretical results are only a first step and come with important limitations. We touch on two of the most important. Incentive-compatible fair representations have substantial information requirements, in particular about market incentives. Our analysis assumes perfect information, which seems impractical. Weakening the information requirement is an important direction for future research that is not addressed in this thesis. Moreover, our results on the amortization of the cost of fairness consider a theoretical blueprint for a data market that has not yet been implemented in the real-world. Replicating the phenomenon of amortization and investigating its robustness to market conditions across data-market settings, in both the real-world and theory, is an important direction for future research that is also not addressed in this thesis. In the remainder of this introduction, we outline the thesis by summarizing the results in the thesis chapters.

1.1 Can pure machine-learning solutions be applied out-of-the-box in data markets?

The literature on machine-leaning fairness has made important progress and contributions in providing pure machine-learning solutions to a number of fairness problems. Pure machine-learning solutions are formulated to be general and applicable across a wide range of domains including data markets, online platforms, and social networks. This is achieved by formulating the problem and solution to include the common structure of the machine-learning problem that occurs across all of the domains while excluding domain-specific structure that occur in only some of the domains. A pure machine-learning solution can then be applied as long as the common structure transfers, but important domain-specific structure may still affect its performance, i.e., the solution

may not be suitable under certain circumstances. To fully understand the applicability of a pure machine-learning solution in a particular domain requires identifying the conditions under which the solutions is suitable as well as the conditions under which the solution is less suitable and may need to be adjusted. In Chapter 2 we study the application of an important pure machine-learning solution, fair representations, in data markets. Our results uncover two broad and complementary themes for understanding the application of pure machine-learning in data markets more broadly.

Data markets complicate the out-of-the-box application of pure machine-learning solutions. We highlight three complications. First, we find that pure machine-learning solutions can be in tension with desiderata that are especially important in data-markets. In particular, we show that fair representations have unfavorable scale and composition properties in data markets.

One of the fundamental and valuable properties of data is that it can often be combined with other data to achieve even greater predictiveness. Combining different data is an important form of composition that can allow greater economic value to be unlocked. We show, unfortunately, that combination can compromise perhaps the most important property of fair representations, an information-theoretic security guarantee against unfairness. And this has consequences for a second important property of fair representations.

Fair representations are flexible, for given input data, there are multiple fair representations, i.e., multiple representations of the data that achieve the security guarantee. Initially, one might imagine that a seller could take advantage of this flexibility to sell different fair representations to different buyers. But, to preserve the security guarantee, the seller must forego combination and commit to selling only one fair representation. This commitment in turn impacts the ability of fair representations to scale in data markets.

We show conditions under which fair representations scale extremely poorly: the commitment to a single fair representation can impose a severe inefficiency that is in addition to and independent of the cost of fairness, i.e., the minimum inefficiency necessary to achieve fairness. Thus, tension between pure machine-learning solutions and data-market desiderata reveal a second complication:

pure machine-learning solutions can be inefficient in data markets—more inefficient than necessary to achieve fairness.

A third complication shows that the issues are not limited to implementation and efficiency; pure machine-learning solutions can raise new fairness concerns even as they solve others. Our results elucidate the flexibility of fair representations, there are many possibilities, yet the seller must pick and commit to one. Our analyses indicate that there can be two or more fair representations for which prediction is equally accurate in the aggregate, but that assign some particular individual different optimal predictions. The seller must therefore potentially deal with fairness concerns related to *arbitrariness* [39]. If an individual receives an unfavorable outcome under one fair representation but not another, and aggregate accuracy does not commend one over the other, how is an unfavorable outcome to this individual to be justified?[3]

Data markets provide opportunity to build on pure machine-learning solutions. When we study the complications that arise from the out-of-the-box application of pure machine-learning solutions, we can identify their sources. We show that these sources can be turned into opportunities to reformulate data-market aware variants of the pure machine-learning solutions that may be able to overcome the complications. We identified the information-theoretic security guarantee as the fundamental source of the complications we discussed. This security guarantee is formulated as a response to a fundamental assumption: that the buyers are malicious adversaries who only want to discriminate and do not care about anything else—including their own utility. Such adversaries do exist and it is important to provide solutions to address them, but it may not be appropriate to model every agent as a malicious adversary all the time and in every context. In economic contexts, agents are often modeled as rational utility-maximizers. We show that fair representations can be generalized using this model of agent behavior. And that this generalization can sometimes recover the inefficiency and achieve fairness. Thus, generalizing a pure machine-learning solution may overcome some complications in data markets and possibly expand their applicability.

[3]This is not unqiue to data markets, and holds for the application of fair representations in any domain. But data markets potentially add a new dimension to the problem of arbitrariness in that a seller may have to address arbitrariness concerns across different buyer's machine-learning problems.

1.2 How do market conditions impact fairness interventions?

In Chapter 2 we studied the question: Do pure machine-learning solutions transpose well in data markets? We identified complications that arise from an out-of-the-box application of fair representations. And we made progress in addressing the complications that we identified by formulating a market-aware variant of fair representations that incorporated more market information. Yet, our investigation did not take into account a fundamental characteristic of markets: markets are dynamic. As market conditions change, market outcomes change. In Chapter 3 we study whether the dynamism of data markets has implications for fairness interventions.

We revisit an existing model of a data market [26]. We capture the dynamism of data markets by formulating a variant that endogenizes data and admits fairness assessment and intervention. Thus, data production at equilibrium in our model depends on market conditions; we focus on market size and the fairness intervention. We analyze the impact of a fairness intervention in our stylized model. We find that market conditions are essential for understanding the impacts of a fairness intervention. Our results uncover two broad and complementary themes for understanding fairness interventions in dynamic data markets.

Neglecting to account for market conditions can severely understate the risk of a fairness intervention. Our analyses indicate that fairness interventions in data markets can also run the risk of backfiring. Under some market conditions, data is produced and transactions occur in the absence of a fairness intervention, but when the intervention is implemented, it can render participation irrational for the sellers so that no data is produced and no transactions occur. This *backfire risk* differs qualitatively from the cost of fairness. The cost of fairness is the loss of efficiency that is incurred in exchange for a fairness benefit. But when an intervention backfires, the cost of fairness is maximal—every agent's full utility—and there is no fairness benefit. Worse still, backfiring can harm the very groups it was intended to benefit. Thus, it is imperative that an intervener identify and mitigate backfire risk. We find that this requires taking market conditions into account: every fairness intervention backfires under some market conditions but not under

others. Without taking market conditions into account, an analysis of intervention risks may miss backfire risk, as well as any other risks that may depend on market conditions. At the same time, we also find that market conditions are important for understanding more than just intervention risks.

Neglecting to account for market conditions can overstate the burden of a fairness intervention. Our analysis of backfire risk indicates that it is critically related to market size. Backfire risk is always mitigated when the data market is sufficiently large. It turns out that market size and how it grows, is also important for understanding the cost of fairness. As one might expect, the cost of fairness grows as the market grows in absolute terms. But perhaps unexpectedly, however, the cost of fairness can shrink as the market grows *in relative terms*. And possibly even astonishingly, the relative cost of fairness can shrink *asymptotically to 0*. Moreover, this can occur simultaneously for *every single agent in the market*. These findings indicate that the overall burden of a fairness intervention, as evaluated in relative terms, can be actually be light. To put this in perspective, consider that there are many real-world market settings in which the burden imposed by transaction costs such as fees, commissions, and taxes is a fixed percentage of every transaction amount. As an example, this means that, the burden imposed by a fairness intervention relative to that of a fixed tax rate can be greater or larger depending on market conditions.

Chapter 2: Incentives Needed for Low-Cost Fair Lateral Data Reuse

2.1 Introduction

It is now well known that there are multiple grounds for moral hazards in the practice of data science (e.g., at data collection, during data cleaning, model specification, at training time, or in subsequent optimizations) [38][1]. Even for the most elementary goal of "fair data-driven algorithms" (statistical parity, see definition below) there are myriad solutions proposed at various stages of data processing. But all those have two assumptions in common: A single entity or administrative domain is in charge of enforcing fairness at all stages, while other participating parties either are fixed or untrusted adversaries. All of those scenarios imply that fair pipelines comes at a substantial operating cost[2]. The issue is further compounded and complicated upon panning out from an individual pipeline, to consider the patterns of sharing, reuse, and consumption of the same published data between separate entities. That is especially pronounced in online targeted advertising, one of the most widespread application of data-driven decisions where it is common for advertisers to aggregate large amounts of data from multiple sources. Consider settings corresponding to one piece of this complex ecosystem: data brokers selling data to advertisers. To ensure fairness in practice, either the data brokers need to sanitize data against an arbitrary advertisers' potential demographic bias, with dire consequences on profit. Or alternatively, in an unregulated market, the advertisers face a dilemma, either incur greater costs to be fair, or sacrifice fairness to increase profit [42]. Little progress has been made in addressing the problem years after evidence that skewed online ads reduce exposure to high earning job for female, limit housing options for some ethnic groups, and is a barrier to career re-entry for older workers [43, 44, 45, 46,

[1] This chapter is a reproduction of [73] without the abstract. The version of record can be accessed at `https://doi.org/10.1145/3412815.3416890`.

[2] The rare exceptions to that rule can be traced to lack of calibration in the training data and model specification. We ignore those cases where fairness essentially counteracts overtraining.

47, 48].

We suggest a fresh new start on achieving fairness in data pipeline, one that departs from the assumptions that the problem is addressed by a single actor through heavy-handed regulation (e.g., the ad-platform, the advertiser, the credit scoring agency, the firm hiring, the firm developing the AI). We formulate for the first time the *incentivizing fairness* problem, inspired but not limited to online advertising as a motivating example. Our single most important assumption is that the entity in charge of the data pipeline faces profit-seeking adversaries: That is a participating entity (e.g., the advertisers of a ad campaign) whose only goal is to maximize profit irrespective of its fairness consequences. This assumption, while common elsewhere (e.g. Game Theory, Economics, Mechanism Design), substantively differs from those made in fairness and fair-representation literature [40]. Our choice rules out, for instance, to work with firms that are actively leveraging data to run an unfair ad campaign *at any cost* (aiming at complete discrimination), or share data with malicious parties. This also requires to make some assumption or have information (however minimal) about how the firm makes profit from the ads. Since our aim is in motivating further exploration of that alternative approach to fairness, the main question we address is "What is the cost of incentivizing fairness in a data pipeline?" "How does it compare to the traditional adversary models?" "Can fairness be made incentive compatible under some simple data manipulation in the pipeline, keeping the design relatively robust to dependencies?"

Our model (see Section 3.3 for notations and formal definitions) in a nutshell focuses on one local step (a fork operation) in data pipelines, which already reveals the crucial role of incentives in achieving fariness. This simple fork pipeline includes a data publishing platform (e.g., a data broker) and multiple data consumer firms (e.g., advertisers interested to target particular individuals who use the platform). Data consumers firms may be a very large number, they are all profit seeking while the publishing platform, which possesses a large database of inviduals it services, has a mandate to achieve statistical parity in outcome. That implies that the publishing platform would only release data if *every* data-consuming firm would in the end select a subset that contains the same fraction of consumers from a given subgroup than in the whole population. If this

model represents hiring ads on Facebook, this objective could be a way to ensure that an advertiser constructs a demographically-balanced custom audience, thereby proportionately targeting female, middle aged or non-white individuals. Profit made by data consumer firms grows in proportion of the accuracy of the classification tasks they perform, just like it would if each ads costs a nominal amount to show but potentially generate a (higher) amount when it reaches a relevant individual. Note that we do not specify how data about individuals are distributed and relate to the various classification tasks, the subset a data consumer firm choses can be rather complex. Features like "custom audience", available on Facebook's ad-platform and others, allow today's online advertisers to make such a selection. Most importantly, the subset and utility derived depend not only on the raw data but on the representation of the data that the platform decides to publish.

This model, however simple, already highlights multiple ways to achieve fairness in that specific interaction. First, it is a perfectly sensible solution to publish data to all consumer firms in a *sanitized* version that keeps demographic features hidden, even from data inference, so they remain secret and discrimination is made impossible. This is in fact the approach advocated in [40] and it forms a natural benchmark, a lower bound on profits. One merit often used to justify this approach is that sharing and reusing this data among consumer firms, and even new ones, creates no additional concern. On the opposite side of the spectrum, one could assume that every consuming firm would first communicate the revenue predicted from each individual in the database, and then it would *delegate* to the publishing platform the selection, where the latter computes among all fair subsets, the one that attains the maximum profit. While this delegation is unpractical, it provides an upper bound of attainable profit. Crucially our model leaves room for a third option: providing data to consuming firms so that fairness is incentive compatible. This holds if choosing a subset based on the published information never results in an unfair subset maximizing profit. Note here that the same data is published for all consuming firms in that fork operation to reuse, possibly in coordination among themselves (what we call *lateral* data reuse). Further data reuse, however, could create unfairness since a new consuming firm, with a very different objective, could possibly select an unfair subset if it accesses this data. Data consuming firms would then face a choice

between being *de facto* fair, or losing revenue. But would that actually lead to different data being published, and more profit?

The main merit of that model, and the result of this paper, is to reveal for the first time that incentive compatible fairness can be a low-cost effective approach:

- We first analyze the cost of using *sanitized* version of the data, formally defined as those achieving demographic secrecy. Multiple solutions in the literature based on calibration of scores or clustering into representative bins have been proposed and evaluated to that effect. It provides individuals in the data with a special protection (i.e., their demographic information cannot be inferred by consuming firms) and automatically translate into some forms of downstream fairness. But we show that evaluating the cost of demographic secrecy, which is specifically distinct from the cost of fairness, reveals a simple but important truth: demographic secrecy may be cost effective for a single data consuming firm, but much more costly when multiple consuming firms are using the same published data. (Section 2.4)

- Given that the costs of fairness and demographic secrecy are only the same in a simple case (a single consuming firm), how large can the gap be in a simple model of individuals' data? And more importantly, can some representations of the data make fairness incentive compatible and recover some of this additional cost? We show the high potential of leveraging incentive compatibility for fairness in the following set of results: While the cost of fairness is linear in the number of firms the added cost of demographic secrecy is exponential, and with high probability fairness can be achieved using incentives with *no* extra cost. Moreover, while this result obviously is a reflection of the data model we assume, it is found for the simplest (independent classification tasks), which makes it likely that this theoretical gaps translate into practical gain. (Section 2.5).

- The results presented above are encouraging, especially because fairness is often considered prohibitive while we clarify that, in simple cases, only demographic secrecy is. It would be premature, even misleading, to conclude that fairness can always be achieved using in-

centives at no extra cost. Relying on incentives to accomplish fairness with data reuse also creates new concerns. We clarify the potential and limitations as we review the potential for such results to generalize and how they motivate new directions in data pipelines. (Section 2.6)

Before presenting the contribution in the order above, we quickly review related work on fair representations and the associated costs they introduce.

2.2 Related Work

Our work is situated in the literature on *fair representations* initiated by Zemel et al. in [40], where the authors consider a setting in which a trusted platform releases data to a single third party. Later work extended the setting to multiple third parties [49], and this is the setting in which we develop our model. A defining feature of fair representations to date has been that demographic information is obfuscated, ideally in an information theoretic manner. In contrast to such *demographically-secret* fair representations, the notion of *incentive-compatible* fair representations that we propose generalizes the notion of a fair representation.

The majority of work on fair representations has focused primarily on the problem of finding a *transformation* of the original dataset that results in a demographically-secret fair representation while preserving as much non-demographic information as possible. Zemel et al. propose an approach based on a discriminative clustering model [40]. Feldman et al. propose an approach that learns a transport map from each group's distribution to the aggregate empirical distribution of the data [50]. Johndrow and Lum generalize this to a statistical model-based approach capable of handling discrete features and an arbitrary choice of target distribution [51]. A number of papers have considered approaches based on adversarial learning with variations in the choice of generator, adversary, and their respective optimization objectives [52, 53, 54, 49]. Such diversity in the details of the form of the raw data, the choice of learning algorithm, the specification of the transformations, and the form of the representation present challenges to theoretical studies of fair representations. We overcome these challenges by focusing on the *computational links* that

transformations create between initially distinguishable individuals by mapping them to the same value.

The fair machine-learning community has identified a need to theoretically study the properties of fair representations [55], although there has been a limited amount of work to date. McNamara, Ong, and Williamson assume that one can measure the distance between the raw datum and the transformed datum, and show how to prove that a fair representation will be demographically secret and how to bound the loss in utility of the resulting representation [56]. In contrast, our model makes no assumptions about the form of, or relationship between the input and output of a fair representation.

A key contribution of this work is to formalize and quantify for the first time the cost of demographic secrecy. This is very closely related to the extensively-studied cost of fairness [37, 57]. Crucially, the cost of demographic secrecy is distinct, and, as we will show in a simple model, sometimes much larger. In particular, the cost of fairness derives from differences in the group-specific statistics, whereas the cost of demographic secrecy derives from computational links between individuals necessarily created to obfuscate demographic information.

2.3 Utility of Fair Representations

The Publisher, Individuals, and Groups A *publisher* has a dataset, perhaps of city hotline phone calls or medical histories. Each datum contains information associated with some individual. Naturally the form, content, and semantics of the data can vary considerably: an individual's Facebook likes, high school transcripts, or ultrasound images from her most recent prenatal visit. We abstract away these details by focusing on the individuals whom we model as elements v of a finite set V. We assume that the individuals in V are distinguishable, that is, associated with a unique datum. Additionally, each individual exclusively belongs to some group g in a finite set of groups G given by a group-membership function $\gamma : V \mapsto G$.

We will often focus on all the members of a group. For each group $g \in G$, we denote by V_g the set of all individuals that belong to g in V, that is, $V_g = \{v \in V : \gamma(v) = g)\}$. We mostly

consider the important special case where there are exactly two groups of equal size: $|G| = 2$ and for $g, g' \in G$, $|V_g| = |V_{g'}|$. We shall say that such V is a *binary-balanced* set.

Transformations and Representations The publisher receives an initial dataset in some form, but is not required to publish the raw data. In particular, the publisher may choose to apply a *transformation*. The details of the transformation can vary considerably as to its purpose, effects, computational complexity, and so on. Perhaps to protect privacy, the values of some data fields are collapsed to achieve k-anonymity; or for compression the top-k components are obtained using Principal Component Analysis; or for transparency, the raw dataset is released. We abstract away these details by focusing primarily on the resultant *representation*. While each datum in the initial dataset is associated with a unique individual, each datum in the representation is associated with *one or more* individuals. Therefore, we model a representation Z as a partition of V. Each part $z \in Z$ represents the individuals whose computational fates the transformation links together by mapping them to the same datum. We denote the set of all possible partitions of a set V by $\Pi(V)$. A transformation is thus a function of the form $r : V \mapsto Z$.

Example 2.3.1. *The* identity representation *I is the partition of $|V|$ singleton sets which models the case in which the publisher publishes the raw data. I has the transformation $r_I(v) = \{v\}$.*

Data Consumers and Automated Decision Systems *Data consumers* use the published data to construct *automated decision systems*. We assume that consumers do so independently. The automated decision systems may take many forms: a risk assessment model that outputs an integral-valued score representing a category of recidivism risk; a clustering algorithm for customer segmentation that assigns to each datum in the dataset a cluster identifier. Moreover, the published data may be used as an input directly and indirectly to multiple algorithms: the published data may be directly fed into a representation-learning algorithm, and recommendations may subsequently be made using the learned representation.

We dispense with most of the differences by focusing on their effects. On input a datum from the published data, an automated decision system assigns an outcome. We assume that the

individuals mapped to the same datum by the transformation are indistinguishable and therefore must be assigned the same outcome. We capture this as follows: data consumer i constructs automated decision system $D_i : Z \mapsto O_i$ which maps parts of Z to outcomes in a consumer-specific set of outcomes O_i.

Example 2.3.2. *(Binary Classifier) One of the most common automated decision systems are binary classifiers. Data consumer i constructing a binary classifier has consumer-specific outcome set* $O_i = \{0, 1\}$*, and an automated decision system of the form* $D_i : Z \mapsto \{0, 1\}$.

Data-Consumer Utility Each data consumer chooses an automated decision system based on a consumer-specific *utility* which captures the relation between the assigned outcomes of the automated decision system and the benefit the consumer receives from those assignments. Given a representation Z, data consumer i is constrained to construct an automated decision system D_i from $\mathcal{D}_i^Z = \{D : Z \mapsto O_i\}$, the set of all automated decision systems whose domain is Z and range is O_i. Therefore, we can model each data consumer's utility as a function of the form $u_i : \mathcal{D}_i^Z \mapsto \mathbb{R}_+$.

Example 2.3.3. *(Unit-Additive Binary-Classification Utility) A data consumer i that constructs a binary classifier* D_i *often derives its benefit from the accuracy of the classifier. There is a class-membership function* $f_i : V \mapsto \{0, 1\}$*, and i wishes the classification* $D_i(v)$ *to match the label* $f_i(v)$ *on as many individuals* $v \in V$ *as possible. If each correct classification contributes a constant unit amount of benefit, we may write*

$$u_i(D_i) = \sum_{v \in V} \mathbf{1}\left[D_i(r(v)) = f_i(v)\right] \tag{2.1}$$

Fairness The publisher is concerned with the fairness of the decisions made by the data consumer's automated decision systems; to operationalize this concern requires formalizing fairness. Many sensible definitions have been put forward in the literature[55], and it is not a priori clear how to select the most appropriate one. In this work, we focus on the notion of demographic parity[55], as a prominent definition in the literature; while this does limit somewhat the applicability

of our results, as we shall see, no definition of fairness obviates the fundamental source of the cost of demographic secrecy (i.e. computational links that are unnecessary for achieving fairness). Henceforth, we use the term fairness as a synonym for demographic parity which requires that for any fixed set of outcomes, the distribution of outcomes across groups be the same.

Definition 2.3.1. *(Demographic Parity) Let D_i be the automated decision system constructed by data consumer i which assigns outcomes $o \in O_i$, and r be a transformation.* D_i satisfies demographic parity *if for every outcome $o \in O_i$, and groups $g, g' \in G$ we have*

$$\Pr\left[D_i(r(u)) = o | \gamma(u) = g\right] = \Pr\left[D_i(r(v)) = o | \gamma(v) = g'\right], \tag{2.2}$$

where the randomness in both probabilities is taken over uniform choice of individual in their respective groups and the randomness of the automated decision systems.

Social Welfare The publisher is also concerned with the *social welfare*, the sum of all the consumer-specific utilities. Since the publisher's choice of representation determines the possible automated decision systems the consumers may construct, we model the social welfare as a function of a representation:

$$u(Z) = \sum_{i=1}^{n} \max_{D \in \mathcal{D}_i^Z} u_i(D). \tag{2.3}$$

We now have all the ingredients to formally define our problem.

Definition 2.3.2. *(Fair Representation Problem) Let V be a finite set of individuals with associated set of groups G and group membership function γ. Let there be n data consumers with a collection of utilities $U = \{u_i : i \in [n]\}$. The* (V, U)-Fair Representation Problem *is to output a representation Z such that*

$$u(Z) = \max_{Z' \in \Pi(V)} u(Z'), \tag{2.4}$$

and the automated decision system D_i satisfies demographic parity for every consumer i. We refer to a pair (V, U) as an instance of the fair representation problem.

Note that the fairness constraint is crucial; the problem is otherwise trivial, publish the identity representation. Moreover, even with the constraint, it is clear that the search space is intractably large for a brute force solution. Thus, one can view demographic secrecy as a design decision that both prunes the representation search space and creates the fairness guarantee against a malicious data consumer.

Definition 2.3.3. *(Demographic Secrecy) Let* (V, U) *be an instance of the fair representation problem. A representation* Z *is* demographically secret *if for every* $z \in Z$ *and* $g \in G$ *it holds that*

$$\frac{|z_g|}{|z|} = \frac{|V_g|}{|V|}. \tag{2.5}$$

We denote the set of all demographically-secret representations by $\Xi(V)$.

Our definition of the fair representation problem anticipates greater flexibility in the choice of representation.

Definition 2.3.4. *(Incentive Compatibility) Let* (V, U) *be an instance of the fair representation problem. We say that a representation* Z *is* incentive compatible *if, for every consumer* i*, the following implication holds:*

$$\textit{If } D_i \in \arg\max_{D \in \mathcal{D}_i^Z} u_i(D), \textit{ then } D_i \textit{ satisfies demographic parity.} \tag{2.6}$$

Note that the set of incentive-compatible representations trivially subsumes the set of demographically-secret representations.

2.4 The Cost of Demographic Secrecy

The same property which makes demographic secrecy attractive as a solution concept also makes it a very strong property: every single part z (i.e. every datum output by the transformation) must have the same demographics, and this must match the overall demographics. In view of the fact that naturally occurring data tend to be highly correlated with the demographics with which

algorithmic fairness is concerned, it is natural to ask: Is there a cost to demographic secrecy? First, consider what is achievable.

Definition 2.4.1. *(Demographically-Secret Social Welfare) Let* (V, U) *be an instance of the fair representation problem. The* demographically-secret social welfare*, denoted* $\delta(V, U)$ *is defined to be*

$$\delta(V, U) = \max_{Z \in \Xi(V)} u(Z). \tag{2.7}$$

We will often simply write δ*.*

We note, that the definition is a scalar value; to realize this value the publisher faces an additional computational problem of finding some representation Z such that $u(Z) = \delta$. In such case we say that *the representation* Z *achieves* δ*.*

The cost of demographic secrecy should be quantified with respect to what each consumer could ideally achieve on their own while being fair.

Definition 2.4.2. *(Fair Social Welfare) Let* (V, U) *be an instance of the fair representation problem, and* $n = |U|$*. The* fair social welfare*, denoted* $\beta(V, U)$ *is defined to be*

$$\beta(V, U) = \sum_{i=1}^{n} \beta_i, \tag{2.8}$$

where

$$\beta_i = \max_{D_i \in D_i^I} u_i(D_i), \tag{2.9}$$

subject to D_i *satisfies demographic parity for every* i*. We will often simply write* β*.*

Definition 2.4.3. *(Cost of Fairness) Let* (V, U) *be an instance of the fair representation problem. The* cost of fairness *is defined to be*

$$CoF(V, U) = u(I) - \beta(V, U). \tag{2.10}$$

Definition 2.4.4. *(Costs of Demographic Secrecy) Let* (V, U) *be an instance of the fair representation problem. The* cost of demographic secrecy *is defined to be*

$$CDS(V, U) = \beta - \delta. \tag{2.11}$$

The relative cost of demographic secrecy *is defined to be*

$$rCDS(V, U) = \frac{CDS(V, U)}{\beta} = \frac{\beta - \delta}{\beta}, \tag{2.12}$$

We will often write CDS and rCDS where the problem instance (V, U) *is clear from the context.*

Note that $\mathrm{CDS}(V, U) \geq 0$. The cost of demographic secrecy is the *minimum* loss in the social welfare that is a consequence of requiring the representation to be demographically secret. A key feature of our model is that the cost of demographic secrecy is *prior* to and *independent* of any cost to the social welfare that results from the information lost in transforming the data into the fair representation. The cost of demographic secrecy places an upper bound on what is achievable by *any* automated decision system constructed by any method. To the best of our knowledge, the literature on fair representations focuses solely on addressing this latter issue, and so it crucially distinguishes our work.

2.4.1 One Data Consuming Firm

Many algorithms for learning fair representations have been developed; to preserve as much relevant information as possible in the fair representation, these algorithms often incorporate an objective term which penalizes loss of predictiveness of a target variable. In privileging one target variable in this way, we can view these algorithms as focusing on the special case of one data consumer.

In evaluation, it is common to compare an automated decision system constructed using the fair representation against one trained on the raw data using an inprocessing fairness intervention. Typically, the evaluation compares the differences in utility and overall fairness achieved. It is

consistently reported that fair representations perform competitively, despite the ostensible severity of demographic secrecy. Our first result shows that this can be anticipated theoretically. Informally, this is so because the publisher can, in theory, virtually construct an automated decision system for the consumer that achieves β.

Theorem 2.4.1. *Let (V, U) be an instance of the fair representation problem such that there is only one data consumer, $|U| = 1$. Then, there exists a demographically-secret representation Z that achieves β; in other words, $\beta = \delta$.*

Proof. Let $D^* : V \mapsto O$ be an automated decision system in $\mathcal{D}^I$ that achieves β and satisfies demographic parity. Define the representation Z to be the partition of V of $|O|$ parts where each part $z_o \in Z$, $o \in O$, consists in all individuals assigned outcome o by D^*, that is, $z_o = \{v \in V : D^*(v) = o)$. Observe that Z satisfies demographic secrecy since D^* satisfies demographic parity. Moreover, given Z, the data consumer can construct $D : Z \mapsto O$ defined by $D(z_o) = o$, and so $u(D) = u(D^*) = \beta$. □

Example 2.4.1. *(College Admissions) As a concrete example, suppose a university is deciding which prospective students to admit from a pool of applicants drawn from two groups, and that the college uses an automated decision system to decide which students to admit. There are data on the students in the form of a score which distills the college's evaluation of the student's ability to thrive and contribute and represents a student's contribution to the college's utility. Yet, for whatever reason, the distributions of student scores differ between the groups. Here we can consider the college as both the publisher and single data consumer, having access to both the raw scores and constructing an automated decision system. When the college admits a ρ-fraction of applicants, it achieves the maximum utility possible while being fair by admitting the top ρ-fraction of applicants from each group. Let $s(v)$ denote the score of individual v, and F_0 and F_1 be the score distribution functions for members of group 0 and 1, respectively. Then, the following transformation results in a demographically-secret representation that achieves the maximum fair utility for the college, $r : V \mapsto [0, 1]$ defined by $r(v) = F_{\gamma(v)}(s(v))$.*

Observe that Theorem 2.4.1 is independent of the data consumer's utility since the critical factor is the assignment of outcomes to individuals. As we investigate the case of multiple data consumers and in the remainder of this paper, we will assume that all the data consumers have unit-additive binary-classification utilities. Although this is a strong and limiting assumption, we feel that it is sensible given the ubiquity of binary classification in data science. Moreover, as with definitions of fairness, no choice of utility obviates the fundamental source of demographic secrecy.

2.4.2 Multiple Data Consuming Firms

Theorem 2.4.1 has the happy consequence that when there is one data consumer, the publisher does not have to worry about any theoretical gap between β and δ. Unhappily, this does not hold generally. Once multiple data consumers uses the same representation, demographic secrecy may come at a cost that is in addition to the cost of fairness.

Theorem 2.4.2. *There exist instances of the fair representation problem* (V, U)*,* $|U| = 2$ *such that* $CDS(V, U) > 0$*.*

Proof. Consider the following problem instance: $V = \{w, x, y, z\}$, and $G = \{0, 1\}$. We have $V_0 = \{w, x\}$, and $V_1 = \{y, z\}$. There are two data consumers with utilities $U = \{u_1, u_2\}$. Both u_1 and u_2 are unit-additive binary-classification utilities corresponding to class-membership functions f_1 and f_2, respectively, specified in Table 2.1. Observe that there are two possible demographically-secret representations. Either $\{\{w, y\}, \{x, z\}\}$ or $\{\{w, z\}, \{y, x\}\}$. In both cases, both pairs disagree on exactly one of the class-membership functions, so any automated decision system constructible from a demographically-secret representation of V must make a mistake on exactly two individuals. Hence $\delta = 6$. On the other hand, given the identity representation, the data consumers could perfectly classify all the individuals and be fair. Thus, $\beta = 8$, and

$$\mathrm{CDS}(V, U) > 0. \tag{2.13}$$

□

Table 2.1: Two class-membership functions f_1 and f_2 and the group-membership function γ over a set of 4 individuals.

	$f_1(\cdot)$	$f_2(\cdot)$	$\gamma(\cdot)$
w	1	1	0
x	0	0	0
y	1	0	1
z	0	1	1

The relative cost of demographic secrecy in the example given in Theorem 2.4.2 is $1/4$. Observe two features of the proof. Every pair of individuals disagree on at least one of the class-membership functions and each class-membership function is fair, assigning in each group the same number of individuals to each class. By a careful construction of a binary error-correcting code, it is possible to construct class-membership functions that scale these properties to binary-balanced sets V of any size.

Theorem 2.4.3. *Let V be a binary-balanced set of individuals, then there exists a U such that*

$$rCDS(V, U) \geq \frac{1}{8}$$

And for many binary-balanced V the relative cost of demographic secrecy can be even more severe.

Corollary 2.4.1. *Let V be a binary-balanced set of individuals such that $|V| = 2^k$, $k \in \mathbb{N}$, then there exists a U such that*

$$rCDS(V, U) \geq \frac{1}{4}. \tag{2.14}$$

The key observation is that demographic secrecy forces a consumer i to misclassify at least one individual whenever a pair of individuals u, v are linked in the representation and their labels differ, $f_i(u) \neq f_i(v)$. If the number of these forced mistakes exceeds the cost of fairness, the number necessary to achieve demographic parity, then those further mistakes are an additional cost. However, although limited by demographic secrecy, the publisher clearly has a lot of power

in whom to link in the representation, and this makes it difficult to prove a statement over the set of all demographically-secret representations. But at least in the case of a binary-balanced set V, the task is made significantly more tractable by observing that we can restrict our attention to a subset of demographically-secret.

Definition 2.4.5. *(Pairing Representations) Let V be a finite set of individuals and Z be a representation of V. We say that Z is a* pairing representation *if for every $z \in Z$ we have*

$$|z| = 2. \tag{2.15}$$

In a binary-balanced set V, exactly half the individuals in any part z of a demographically-secret representation Z must belong to one group, and the rest to the other. Thus, splitting every z into $|z|/2$ parts with one member of each group is a demographically-secret pairing representation Z'. So any automated decision system that can be constructed using Z can be perfectly simulated using Z'. Therefore, we need only focus on the set of demographically-secret pairing representations.

If we can guarantee that every pair of individuals disagree on at least some number of class-membership functions, then that would provide a bound on CDS from below. Given n class-membership functions, there is a natural mapping $b : V \mapsto \{0,1\}^n$ from individuals to binary strings defined in terms of the class-membership functions as $b(v)_i = f_i(v)$. Note that the number of class-membership functions on which individuals u, v disagree is the Hamming distance between their respective binary strings $d_H(u, v)$. Alternatively, if we have a collection of $|V|$ binary strings of dimension n with some lower bound k on their Hamming distances, then we can construct n class-membership functions over V with the property that every pair of individuals disagrees on at least k labels.

If we can construct an arbitrarily large collection of binary strings with a large lower bound on their pairwise minimum Hamming distance, we can ensure that any pairing representation will force the consumers to misclassify many individuals across all the class-membership functions. Coding Theory guarantees that such collections of binary strings exists, and gives algorithms for

producing them. However, in application to the setting of fairness considered in this paper, we are faced with a novel challenge for the construction of binary codes: we must also show that the cost of fairness induced by the assignment of strings to individuals is not large.

Definition 2.4.6. *(Fair Codes) Let $K \subseteq \{0,1\}^n$ be a set of m binary strings. We say that K* is a (n,k,m)-fair code *if the following properties holds: For every $p, q \in K$, $p \neq q$ we have*

$$d_H(p,q) \geq k, \tag{2.16}$$

and there exists a partition P of K into two equal sized sets S and T such that for every i,

$$\sum_{s \in S} s_i = \sum_{t \in T} t_i. \tag{2.17}$$

We call P a fair partition of K.

Observe, that given a binary-balanced set V of size m and an (n,k,m)-fair code K, using a fair partition of K, one can construct class-membership functions over V which satisfy demographic parity, and for which there is therefore no cost of fairness. Given an $(2n,n,m)$-fair code K, we can construct a $(4n,2n,2m)$-fair code K' proceeding in the following manner. For each $k \in K$, construct $j \in \{0,1\}^{4n}$ by:

$$j_{2i}j_{2i+1} = \begin{cases} 00 & k_i = 0 \\ 11 & otherwise \end{cases}$$

Let J be the set of all strings so derived. For each $k \in K$, construct $q \in \{0,1\}^{4n}$ as follows:

$$q_{2i}q_{2i+1} = \begin{cases} 01 & k_i = 0 \\ 10 & otherwise \end{cases}$$

Let Q be the set of all strings derived. Define $K' = J \cup Q$. By construction, $J \cap Q = \emptyset$ so

$$|K'| = |J| + |Q| = 2|K|.$$

Observe that every $p, q \in K'$, $p \neq q$ are at a distance at least $2n$ apart. Moreover, the set $S' \subseteq K'$ of strings constructed from S and the set $T' \subseteq K'$ of strings constructed from T form a fair partition of K'. So K' is a $(4n, 2n, 2m)$-fair code. Using this construction, if we start from the set $K = \{00, 01, 10, 11\}$ with fair partition $\{S, T\}$, $S = \{00, 11\}$ and $T = \{01, 10\}$ we can obtain a $(2^{k-1}, 2^{k-2}, 2^k)$-fair code K. We also note that this construction will result in strings that are at exactly a distance 2^{k-2}. We have proved the following lemma:

Lemma 2.4.1. *($(2^{k-1}, 2^{k-2}, 2^k)$-Fair Codes) For every $k \geq 1$, there exists a $(2^{k-1}, 2^{k-2}, 2^k)$-fair code K, and there exist $p, q \in K$, such that $d_H(p, q) = 2^{k-2}$.*

We now prove of Theorem 2.4.3.

Proof. Let $k = \lfloor \log |V| \rfloor$ and $S \subseteq V$ be a binary-balanced subset of V of size 2^k. Let K be a $(2^{k-1}, 2^{k-2}, 2^k)$-fair code. Use K to assign strings to individuals in S so that the resulting class-membership functions are fair over S. We still need to assign labels to all the individuals in $V \setminus S$. Clearly, if we assign them all the same labels on every function, that is $f_i(u) = f_i(v)$ for every $i \in [2^{k-1}]$ and $u, v \in V \setminus S$, then all n class-membership functions f_i will satisfy demographic parity over all of V. We must find a suitable class-membership. Let $x, y \in K$ such that $d_H(x, y) = 2^{k-2}$. Consider the string $z \in \{0, 1\}^{2^{k-1}}$ obtained from x by flipping the first 2^{k-3} bits in x on which x and y differ. Observe that for all $w \in K$ we have,

$$d_H(w, x) \leq d_H(w, z) + d_H(z, x), \tag{2.18}$$

and it follows that

$$2^{k-3} \leq d_H(w, z).$$

Assign the string z to every individual in $V \setminus S$. Since the resulting class-membership functions all satisfy demographic parity,

$$\beta = 2^{k-1}|V|$$

Let Z be any demographically-secret pairing-representation of V. Define the sets $A = \{\{u, v\} \in$

$Z : u, v \in S\}$, $B = \{\{u, v\} \in Z : u \in S, v \in V \setminus S\}$, and $C = Z \setminus (A \cup B)$. Clearly,

$$\delta \leq |V|2^{k-1} - \sum_{\{u,v\}\in Z} d_H(u, v).$$

Bound the summation by,

$$\begin{aligned}\sum_{\{u,v\}\in Z} d_H(u, v) &= \sum_{\{u,v\}\in A} d_H(u, v) \\ &+ \sum_{\{u,v\}\in B} d_H(u, v) \\ &+ \sum_{\{u,v\}\in C} d_H(u, v) \\ &\geq |A|2^{k-2} + |B|2^{k-3} \end{aligned} \tag{2.19}$$

Additionally observe that

$$2|A| + |B| = |S|.$$

Solving for $|B|$ and substituting we obtain

$$\sum_{\{u,v\}\in Z} d_H(u, v) \geq |A|2^{k-2} + (|S| - 2|A|)2^{k-3} \geq |S|2^{k-3} \geq |V|2^{k-4},$$

since $|S| \geq V/2$. Therefore,

$$\delta \leq \frac{7}{8}2^{k-1}|V|$$

We conclude,

$$\mathrm{CDS}(V, U) \geq \frac{1}{8}$$

□

If $|V|$ is a power of 2, then the $(2^{k-1}, 2^{k-2}, 2^k)$-fair code provides a string for every individual, so $|B| = 0$. This proves Corollary 2.4.1.

2.5 Gains of Incentivizing Fairness

In the examples presented so far, the functions were fair, so that if the data consumers were rational, then the publisher could release the raw data and the consumers would be fair incidentally in maximizing their utility; the identity representation is a trivial incentive-compatible representation. But the field of fair machine learning is motivated by observed unfairness in real-world data. In this section, we therefore study the following questions: "Do there exist non-trivial incentive-compatible representations?", "How commonly do they exist?", and "How sizable can their gains be?".

We give an answer to these questions by analyzing a simple model for randomly generating fair representation problem instances where every data consumer has a unit-additive binary-classification utility. A straightforward way to do so is to randomly sample an underlying class-membership function for each data consumer. The simplest random process is arguably one which picks each class-membership function by assigning each individual to the positive class with some probability p. Thus, each class-membership function corresponds to one of the data consumers. In what follows, we will mostly elide the difference between the class-membership functions and unit-additive binary-classification utilities and refer to them interchangeably via their natural isomorphism.

Definition 2.5.1. *(Random Functions Model)* The Random Functions Model (RFM) *on input a set of individuals V and parameters $n \in \mathbb{N}$ and $p \in [0, 1]$. RFM outputs a collection of n class-membership functions $\{f_i : V \mapsto \{0, 1\}\} = RFM_p(V)$, sampling function f_i by setting $f_i(v) = 1$ with probability p and 0 otherwise, for each individual $v \in V$. We will often suppress p.*

Unlike the examples we have presented, RFM can output class-membership functions which do not satisfy demographic parity. Thus, in analyzing RFM, we are in the more realistic and interesting regime of datasets where a rational data consumer will not be fair given the raw data. We will first present and discuss our main results.

In fact, when n grows logarithmically in $|V|$ and with a mild condition on p, we can establish

just how unfair a rational data consumer will be in expectation.

Theorem 2.5.1. *Let V be a binary-balanced set, p such that $\frac{2}{|V|} \leq p \leq 1 - \frac{2}{|V|}$, and $U = RFM_p(V)$, then*

$$\mathbb{E}\left[CoF(V,U)\right] = \Theta\left(n\sqrt{|V|}\right). \tag{2.20}$$

For comparison, we can establish the following lower bound on the expected cost of demographic secrecy.

Theorem 2.5.2. *Let V be a binary-balanced set, p such that $\frac{2}{|V|} \leq p \leq 1 - \frac{2}{|V|}$, and $U = RFM_p(V)$, then*

$$\mathbb{E}\left[CDS(V,U)\right] = \Omega\left(2^{n/2}\sqrt{|V|}\right). \tag{2.21}$$

We see that both the expected costs of fairness and demographic secrecy grow polynomially in $|V|$. However, as the expected cost of fairness grows linearly in n, the expected cost of demographic secrecy grows *at least* exponentially! Amazingly, we can also show that with high probability, an incentive-compatible representation will exist that can recover not just some, but *all* of the cost of demographic secrecy.

Theorem 2.5.3. *Let V be a binary-balanced set, $|V| \geq 2^{20}$, $n = \frac{1}{4}\log|V|$, p such that $\frac{2}{|V|} \leq p \leq 1 - \frac{2}{|V|}$, and $U = RFM_p(V,U)$, Then with probability at least $7/10$, it will be possible to construct an incentive-compatible representation that achieves β.*

A crucial issue in our main results is the choice of n.[3] We argue that $n = O(\log|V|)$ is a sensible and interesting choice. Consider our motivating example of targeted online advertising, constant n would correspond to a constant number of advertisers, regardless of the number of people in the advertising audience, which does not seem to us realistic.

So n should grow in $|V|$, in which case we must ask of what order? We take polynomial growth as the upper limit as it would seem that any one individual may economically sustain at most a constant number of advertisers, so that $n = O(|V|)$. While we do not present the analysis here,

[3]Of course, in practice, a publisher may not have the power to set n. We do so here purely for the purposes of theoretical analysis to obtain qualitative results.

one can show that, for n that is polynomial (in particular, including sublinear) in $|V|$, the situation with the relative cost of demographic secrecy is asymptotically equivalent to that of the constructed examples in Theorem 2.4.3 (i.e. asymptotically constant).

We therefore focus on logarithmic growth, where the dynamics are more subtle. Moreover, logarithmic growth arguably captures the slowest reasonable order of growth in practice, increasing the scope of the implications of our results. We now turn to proving the main results. In our analyses, we assume that V is a binary-balanced set for ease of presentation. They can be extended to binary-unbalanced sets to obtain qualitatively the same results.

The following lemma will prove extremely helpful.

Lemma 2.5.1. *Let X_0 and X_1 be independent, identically distributed binomial random variables with parameters n and p such that $1/n \leq p \leq 1 - (1/n)$, then*

$$\mathbb{E}\left[|X_0 - X_1|\right] = \Theta\left(\sqrt{Var\left[X_0\right]}\right). \tag{2.22}$$

Proof. We first show that

$$\mathbb{E}\left[|X_0 - X_1|\right] = \Omega\left(\sqrt{\mathrm{Var}\left[X_0\right]}\right). \tag{2.23}$$

Let ξ be the event that $X_0 > \mathbb{E}\left[X_0\right]$ and $\mathbb{E}\left[X_1\right] \geq X_1$. Conditioning on ξ we have

$$\mathbb{E}\left[|X_0 - X_1| \,|\xi\right] = \mathbb{E}\left[|X_0 - \mathbb{E}\left[X_0\right] + \mathbb{E}\left[X_1\right] - X_1| \,|\xi\right] \tag{2.24}$$

$$= \mathbb{E}\left[X_0 - \mathbb{E}\left[X_0\right]|\xi\right] + \mathbb{E}\left[\mathbb{E}\left[X_1\right] - X_1|\xi\right]. \tag{2.25}$$

Observe that

$$\mathbb{E}\left[X_0 - \mathbb{E}\left[X_0\right]|\xi\right] = \mathbb{E}\left[X_0 - \mathbb{E}\left[X_0\right]|X_0 > \mathbb{E}\left[X_0\right]\right], \tag{2.26}$$

and

$$\mathbb{E}\left[\mathbb{E}\left[X_1\right] - X_1|\xi\right] = \mathbb{E}\left[\mathbb{E}\left[X_1\right] - X_1|\mathbb{E}\left[X_1\right] \geq X_1\right]. \tag{2.27}$$

Since X_0 and X_1 are independent and identically distributed we have

$$\mathbb{E}\left[\mathbb{E}\left[X_1\right] - X_1 | \mathbb{E}\left[X_1\right] \geq X_1\right] = \mathbb{E}\left[\mathbb{E}\left[X_0\right] - X_0 | \mathbb{E}\left[X_0\right] \geq X_0\right]. \tag{2.28}$$

Therefore

$$\mathbb{E}\left[|X_0 - X_1|\right] \geq \mathbb{E}\left[|X_0 - X_1| \, |\xi\right] \geq c\mathbb{E}\left[|X_0 - \mathbb{E}\left[X_0\right]|\right], \tag{2.29}$$

for some constant c which depends on n and p. By assumption, we have that $1/n \leq p \leq 1 - (1/n)$; applying the Berend-Kontorovich Inequality (see Appendix A.0.1) we obtain

$$\mathbb{E}\left[|X_0 - \mathbb{E}\left[X_0\right]|\right] \geq \sqrt{\frac{\mathrm{Var}\left[X_0\right]}{2}}. \tag{2.30}$$

And therefore

$$\mathbb{E}\left[|X_0 - \mathbb{E}\left[X_0\right]|\right] = \Omega\left(\sqrt{\mathrm{Var}\left[X_0\right]}\right). \tag{2.31}$$

We now show that

$$\mathbb{E}\left[|X_0 - X_1|\right] = O\left(\sqrt{\mathrm{Var}\left[X_0\right]}\right), \tag{2.32}$$

which follows readily since

$$\mathbb{E}\left[|X_0 - X_1|\right] \leq \sqrt{\mathrm{Var}\left[X_0 - X_1\right]} = \sqrt{2\mathrm{Var}\left[X_0\right]}. \tag{2.33}$$

This completes the proof. □

For balanced-binary sets the cost of fairness of a single function is simply how unfair it is in the following sense.

Definition 2.5.2. *(Function Disparity) Let V be a binary-balanced set, and $f : V \mapsto \{0, 1\}$ be a binary-valued function over V. The* disparity of f *is defined to be*

$$\epsilon(f) = \sum_{u \in V_0} f(u) - \sum_{v \in V_1} f(v). \tag{2.34}$$

Since the disparity of any function is clearly the difference of two binomial random variables, and the functions output by RFM are independent and identically distributed, applying Lemma 2.5.1 establishes Theorem 2.5.1.

We now turn to proving Theorem 2.5.2. The cost of demographic secrecy depends ultimately on the functions output by RFM. Analyzing a fixed output of RFM with respect to the cost of demographic secrecy seems hard; we need a proxy. Observe that the parts of a demographically secret representation create computational links between collections of individuals. For a given part, every function on which a pair of individuals in the part disagree enforces at least one mistake that any binary classifier must make in classifying that part. We can formalize this quantity as follows.

Definition 2.5.3. *(Cost of a Representation) Let V be a set of individuals, Z be a representation of V, and U be a collection of n unit-additive binary classification utilities.* The cost of the representation Z *is defined to be*

$$c(Z) = \sum_{z \in Z} \sum_{u,v \in z} \sum_{i \in [n]} \mathbf{1}\left[f_i(u) \neq f_i(v)\right]. \tag{2.35}$$

Now we would like to bound, probabilistically, the cost of any demographically-secret representation Z from below. Each individual u is collectively assigned a binary string, or code, $\ell = k(u)$ by the functions output by RFM. For each code ℓ, the functions will assign some number $m_{0,\ell}$ of individuals in group 0, and some number $m_{1,\ell}$ of individuals in group 1 code ℓ. Denote the difference by $\epsilon(\ell) = m_{0,\ell} - m_{1,\ell}$ and call this the code difference. Observe that in a binary-balanced set, for each code ℓ, *any* demographically-secret representation must pair at least $|\epsilon(\ell)|$ individuals with code ℓ out-of-code. These individuals will contribute at least 1 to the cost of the representation. Finally, summing over all absolute code differences counts each individual twice, so the cost of any demographically-secret representation is at least one half the sum of absolute code differences. We have proved the following theorem.

Theorem 2.5.4. *Let V be a binary-balanced set, and U be a collection of n unit-additive binary-*

classification utilities. We have that for any demographically-secret representation Z of V,

$$c(Z) \geq \frac{1}{2} \sum_{\ell \in L} |\epsilon(\ell)|, \tag{2.36}$$

where $L = \{0,1\}^n$.

Fortunately, we can productively analyze the sum of absolute code differences.

Lemma 2.5.2. *Let V be a binary-balanced set, p such that $\frac{2}{|V|} \leq p \leq 1 - \frac{2}{|V|}$, $U = RFM_p(V)$ be a collection of n functions, and $L = \{0,1\}^n$, then the expected sum of absolute code differences is*

$$\mathbb{E}\left[\sum_{\ell \in L} |\epsilon(\ell)|\right] = \Theta\left(2^{n/2}\sqrt{|V|}\right), \tag{2.37}$$

Proof. For each $g \in G$, and $\ell \in L$, define the random vector $X_g \in \mathbb{R}^{|L|}$,

$$X_{g,\ell} \coloneqq \sum_{u \in V_g} \mathbf{1}\left[\ell = k(u)\right], \tag{2.38}$$

and

$$Y \coloneqq X_0 - X_1, \tag{2.39}$$

so that

$$\|Y\|_1 = \sum_{\ell \in L} |\epsilon(\ell)|. \tag{2.40}$$

Write

$$\mathbb{E}\left[\|Y\|_1\right] = \mathbb{E}\left[\sum_{\ell \in L} |Y_\ell|\right] = 2^n \mathbb{E}\left[\left|X_{0,\ell} - X_{1,\ell}\right|\right]. \tag{2.41}$$

Note that $X_{0,\ell}$ is a binomial random variable with parameters $|V|/2$ and 2^{-n}, and apply Lemma 2.5.1 to obtain,

$$\mathbb{E}\left[\|Y\|_1\right] = 2^n \Theta\left(\sqrt{\frac{|V|}{2}\frac{1}{2^n}\left(1 - \frac{1}{2^n}\right)}\right) = \Theta\left(2^{n/2}\sqrt{|V|}\right). \tag{2.42}$$

□

Theorem 2.5.2 follows as a consequence. Finally, we turn to proving Theorem 2.5.3. It is due to two properties of RFM: First, although most of the random functions output by RFM are unfair, most of the time they will not be too unfair.

Lemma 2.5.3. *Let V be a binary-balanced set, and $\{f_i\}$ be a collection of n functions output by RFM, then with probability at least $9/10$, the absolute disparity of any function $|\epsilon(f_i)|$ will be at most $\sqrt{|V|\ln(40n)}$.*

Proof. For every $g \in G, u \in V_g$, and $i \in [n]$ define random variables

$$X_{g,i,u} \coloneqq f_i(u), \tag{2.43}$$

and

$$X_{g,i} \coloneqq \sum_{u \in V_g} X_{g,i,u}. \tag{2.44}$$

Then

$$\epsilon(f_i) = X_{0,i} - X_{1,i}. \tag{2.45}$$

Apply Hoeffding's Inequality to every $X_{g,i}$ to obtain

$$\Pr\left[\left|X_{g,i} - \mathbb{E}\left[X_{g,i}\right]\right| \geq \sqrt{|V|\ln(40n)/4}\right] \leq \frac{1}{20n}. \tag{2.46}$$

Define indicator random variables

$$Y_{g,i} \coloneqq \begin{cases} 1 & \left|X_{g,i} - \mathbb{E}\left[X_{g,i}\right]\right| \geq \sqrt{|V|\ln(40n)/4} \\ 0 & \text{otherwise.} \end{cases} \tag{2.47}$$

and

$$Y \coloneqq \sum_{g \in G} \sum_{i \in [n]} Y_{g,i}. \tag{2.48}$$

By Markov's Inequality we have

$$\Pr[Y \geq 1] \leq \frac{1}{10} \tag{2.49}$$

and therefore

$$\Pr[Y < 1] \geq \frac{9}{10}, \tag{2.50}$$

which is the event that the absolute deviation of any $X_{g,i}$ about its mean is at most $\sqrt{|V| \ln(40n)/4}$. Since the random variables $X_{g,i}$ are all independent and identically distributed the absolute disparity of any function is at most

$$|\epsilon(f_i)| \leq 2\sqrt{|V| \ln(40n)/4} = \sqrt{|V| \ln(40n)}. \tag{2.51}$$

□

Second, although RFM will often output collections of functions that induce large code differences, in both groups, many individuals will be assigned every code.

Lemma 2.5.4. *Let V be a binary-balanced set, $|V|^4 \geq 12$, and $\{f_i\}$ be a collection of $n = \frac{1}{4}\log|V|$ functions output by RFM. Then with probability at least $8/10$, for every code $\ell \in \{0,1\}^n$, at least*

$$\frac{|V|^{3/4}}{10} \tag{2.52}$$

individuals of each group will have the code.

Proof. Let $X \in \mathbb{R}^{2\times 2^n}$ where $X_{i,j}$ gives the number of individuals in group i assigned code j. The row X_i is a draw from a multinomial distribution with parameters $|V|/2$ and probability vector p, $p_i = 2^{-n}$ for every i. Applying the Bretagnolle-Huber-Carol Inequality (see Appendix A.0.2), we have

$$\Pr\left[\sum_{j=1}^{2^n}\left|X_{0,j} - \frac{|V|}{2\cdot 2^{-n}}\right| \geq 2\sqrt{\frac{|V|}{2}}\lambda\right] \leq 2^{2^n}\exp(-2\lambda^2). \tag{2.53}$$

We desire

$$2^{2^n} \exp(-2\lambda^2) \leq \frac{1}{10} \tag{2.54}$$

Solving for a lower bound on lambda we obtain

$$\lambda \geq \frac{5 \cdot 2^n}{8 \log e}, \tag{2.55}$$

when $|V|^4 \geq 12$. Choosing λ at the lower bound, we have that with probability at least $9/10$, the number of individuals with a given code will differ from its expectation by more than

$$\sqrt{2|V|}\lambda. \tag{2.56}$$

In which case, we can bound the number of individuals with each code from below by,

$$\frac{|V|}{2 \cdot 2^n} - 2\sqrt{\frac{|V|}{2}}\lambda \implies \tag{2.57}$$

$$\frac{|V|^{3/4}}{2} - \frac{5|V|^{3/4}}{8 \log e} > \frac{|V|^{3/4}}{10} \tag{2.58}$$

The same analysis applies to X_1. Consequently, we can bound the probability that this lower bound applies to both groups from below by $8/10$. □

By relaxing the demographic secrecy constraint, an incentive-compatible representation can exploit these properties of RFM to create necessary links that reduce unfairness and avoid creating unnecessary links that impose further costs to the social welfare.

An incentive-compatible representation can close the disparity $|\epsilon(f_i)|$ of the function f_i by pairing individuals from the different groups according to $\epsilon(f_i)$. If $\epsilon(f_i)$ is positive, then more members of group 0 are labeled 1 than members of group 1. By pairing a member of group 0 labeled 1 with a member of group 1 labeled 0, the disparity of f_i is diminished by 1. Constructing $\epsilon(f_i)$ such pairs ensures that a rational consumer would construct an automated decision system that satisfies demographic parity. If $\epsilon(f_i)$ is negative, then the situation is reversed, and an incentive-compatible

representation would have to pair members of group 0 labeled 0 with members of group 1 labeled 1. We call such a pair a disparity-diminishing pair.

If, for every function f_i, it is possible to make $|\epsilon(f_i)|$ disparity-diminishing pairs of individuals, then doing so—and no more—yields an incentive-compatible representation that achieves β. When there are many individuals in both groups assigned to every code, then exactly the necessary number of disparity-diminishing pairs can be made. This is the core of the proof of Theorem 2.5.3, which we now present.

Proof. With probability at least $9/10$, the maximum absolute disparity of any function will be at most $\sqrt{|V| \ln(40n)}$. And with probability at least $8/10$, each group will have at least

$$\frac{|V|^{3/4}}{10} \tag{2.59}$$

individuals with each code. It is straigtforward to check that the inequality is satisfied for $|V| = 2^{20}$. Therefore, these events will occur together with probability at least $7/10$. When they do, for each function f_i, we can make $|\epsilon(f_i)|$ disparity-diminishing pairs using individuals with code x from one group and individuals from the other group with code x', where $x_i \neq x'_i$ in the necessary way, and $x_j = x'_j$ for all $j \neq i$ to construct an incentive-compatible representation which achieves β. □

2.6 Discussion

Our paper proposes a different way to implement fairness in data pipelines. While it is encouraging that our approach radically improves on accuracy costs, it does not come for free. This is why we feel it is important to address, beyond the results aforementioned, the limitations and ramifications of realizing fairness through incentives.

2.6.1 Do accuracy gains generalize?

Demographic secrecy adds a large (exponential) and not-strictly-necessary cost to achieving fairness. Indeed, that cost can entirely be removed by proper incentives. But could that be an arte-

fact of our simplifying assumptions? We offer elements to help inform that important discussion.

- All points to our results generalizing to independent classification tasks with various sensitivity (e.g., one consuming firms looking for a target containing 5% of the nodes, while another targets 80% of them), and to achieve statistical parity a finite number of groups. Therefore, we suspect incentives can keep fairness low cost even with intersectionality (e.g., handling gender and age at the same time).

- The assumption that classification tasks of various data consuming firms are independent seems hardly justified. In practice multiple firms are conducting similar or even identical predictions (regarding credit, or interest in specific purchases) that would correlate. At least, the high gain of incentive trivially generalizes to a scenario where all firms are among n *types* if prediction by different types are sufficiently different to be considered independent, and the number of types n grows with data size beyond $\log(|V|)$. The case with correlated types, and prediction correlating with group membership is more challenging to incentives, making its exploration all the more important in future work.

- Finally, one could dispute our choice of statistical parity as a meaningful accomplishment of fairness, and argue that our results disappear if another fairness goal is used. We have not fully analyzed that aspect, partly because statistical parity is so commonly used, and a consensus is slow to emerge on what to include as practical conditions for fairness. It seems that several other group based definition (based on false positive, equality of opportunity, calibration) would reproduce the same essential tradeoff, while others (individual fairness) can be harder to model from an incentive viewpoint.

We would like to cautiously advise the reader against concluding from our results that incentive compatible fairness *generally* comes at no cost. That result remains surprising, especially when other techniques appear prohibitive. However, we are hopeful that more results can be found (positive or negative) to better appreciate its real potential.

2.6.2 What are impediment and limitations?

Even in a case where our result applies and gains are expected, would achieving fairness through incentive be practical and robust?

- We offered (so far) existential results: concentrate on the potential of incentives to bring fairness at low cost, but not on how that could be implemented. However, our proof highlights the combinatorial flexibility offered by incentives (esp. in comparison with demographic secrecy). That alone suggests to us that it should be possible to regain part of this large gap with suitable data representation. We also feel that answering that question requires to carefully understand how data consuming firms would communicate with the platform. So it moves away from the stylized model we have, and become more application specific. As an example, for online advertising, one can study which decentralized bidding process make fairness incentive compatible, and optimize for accuracy. It could be different if our model is used to study data purchase from public institution.

- Fairness is provided here by anticipation of incentives and it leaves the system vulnerable to some deviation. For instance, the data consuming firm could *in theory* first misrepresents its interest/utility/bids as a way to gain information; once the data are disclosed the firm may follow a different strategy, possibly an unfair one, for a greater profit. Requesting firms to commit to a strategy in advance seems too heavy handed as other solutions exist: For instance, since fairness of outcome is not hard to measure, a firm deviating significantly from it would eventually be noticed. Auditing the firm for a mismatch in anticipated and observed behaviors can deter such misbehavior at lost cost.

Finally, we wish to clarify once again the unavoidable limitation baked in the design: since incentive compatible data release are not strictly bound by demographic secrecy, a data consuming firm can learn demographic or sensitive attributes from it. That firm can share it with a 3rd party which would later reuse it for nefarious purpose. We work under the assumption that this threat is *not* a concern: For instance, the firms accessing the data would not be able to make inference and

share with other party without a considerable risk. That risk can be increased using combinations of data watermarking, internal audits, and regulation. Note that we are not aware of any cost-effective solution to the aforementioned threat: it does require either to restrict data access to a single data firm, to limit data access (with cryptographic primitives) to prevent any data reuse, or to use a demographically-secret representation. All those incur a high operating or accuracy cost. We expect that the cost is so high that many data pipelines would use alternative models like ours.

2.6.3 Data Reuse and Composition

Our results show that demographically-secret representations may be costly for lateral data reuse in which a single dataset is reused across multiple indendent prediction tasks or shared with multiple third parties. Our results also show that the sequential composition property of demographically-secret representations is not robust to aggregation; demographic information may leak when individually demographically-secret representations are combined[4].

We have made direct progress in addressing the former limitation by demonstrating that incentive-compatible representations may recoup some of the cost of demographic secrecy in lateral data reuse; yet it would seem that incentive-compatible representations are otherwise a step backwards for fair composition in that they achieve the utility gains precisely by exposing demographic information. However, note that demographic parity of sequential composition implies the problematic computational links via demographic secrecy since it must hold in the special case of a single automated decision system; therefore, sequential composition is inimical to aggregation.

To provide tools for data scientists to combine multiple datasets and reason about their fairness properties under composition, requires a different approach. Although, strictly speaking, we have not directly studied composition in this work, our results do provide some hope that such an approach may exist and even suggest a possibility. The combinatorial flexibility of fair representations suggest that incentive-compatible representations may be applicable across a wide range of settings diverse in the details of their utilities, definitions of fairness, and patterns of data reuse.

[4]To see this, consider the example given in Theorem 2.4.2, given both possible demographically-secret representations one can not only recover an individual's group, but in fact their very identity.

Further, this flexibility might be amplified by approaching composition with the goal of *controlling unfairness leakage* as opposed to completely preventing it. Can, for example, a principled approach be developed that makes reasonable assumptions on the structure of firms' utilities and behavior as in [47], that allows preserving utility and controlling unfairness?

2.7 Conclusion

Few people today dispute the importance to remove bias and discriminations emerging in applications of machine learning, especially in technical research venues, and regulatory bodies. But the practice of machine learning, involving multiple stages of pipelining and data reuse between interactive parties that cannot be transparently trusted, is rarely introduced in the analysis. The limited tools available today—mostly, relying on demographic secrecy and its downstream invariance—contribute to the perception that providing fairness end-to-end guarantees come at a prohibitive cost. Practitioners often resort to *piecewise fairness*, essentially testing each pipelining step locally on a best effort voluntary basis to identify bias amplification and address it (in the best case), or hide it (in the worst case).

Our results clarified that a part of the currently perceived large cost of fairness in fact serves a narrower purpose: offering a protection against specific malicious data sharing and reuse, that are all strictly speaking *outside* the pipeline. Fairness can sometimes be achieved at a much lower cost when those egregious reuses can be prevented in other ways. We invite relevant research communities to contemplate alternatives where pipelines leverage incentives as a vector to align utility with fairness.

Chapter 3: The Cost of Fair Production in a Data Market

3.1 Introduction

Fairness is now widely recognized as an important system requirement in machine learning. The Biden administration's Blueprint for an AI Bill of Rights enumerates five principles including *Algorithmic Discrimination Protections*: "[y]ou should not face discrimination by algorithms and systems should be used and designed in an equitable way" [58]. However we observe an increasing gap between such stated principles and the everyday practice of integrating machine learning in various pipelines. While many factors contribute to this gap—including the difficulty to agree on a fairness goal and multiple technical implementation challenges—we focus here on a fundamental argument used against fairness intervention: the "cost of fairness," loosely defined as the decrease in accuracy/efficiency that the system incurs when fairness needs to be satisfied, is often too high and could result in undesirable trade-offs such as reduced service or a prohibitive cost to individuals.

One reason why this argument is central, yet not always explicitly accounted for, is the myriad techniques proposed in the area of *fair machine learning* that focus on optimally managing the trade-off between accuracy and fairness at the machine-learning training stage. The literature abounds in either (1) articles proposing new ways to achieve a minimum possible cost of fairness under given fixed fairness constraints [59, 60, 61, 62, 63], or (2) other works finding that the cost remains too high while proposing alternatives that incurs a lower cost of fairness [64, 65, 66, 67]. A fundamental assumption in such works is that the raw data are fixed; an intervention may transform the raw data, for example to hide demographic information [40], but the raw data is the same whether or not an intervention is applied, and no matter the severity of the cost of fairness. In this way, such works place the raw data outside the scope of any fairness–efficiency tradeoff.

But a fairness intervention may impact the extraction of value from the raw data, perhaps through information loss, and little is known about how fairness interacts with the economic incentives to produce and sell data.

For instance, ensuring *demographic balance* (see Definition (3.3.4) below) in training datasets has been shown to bring fairness benefits [68, 42]. Under general conditions, demographic balance translates directly into equal excess errors [64]; this has also been studied in the context of data production in [69]. Using representative data is widely recognized as important for building algorithmic discrimination protections [58] and for testing machine-learning systems for unfairness in practice [68]. Moreover, it can be evaluated based only on the production decisions of the sellers, independently of downstream modeling decision made by data consumers. However, data produced at equilibrium by producers (or "sellers") in an efficient marketplace may deviate from demographic balance to better focus on utility maximizing exchanges. For example, data produced for marketing applications may skew towards economically-advantaged groups. Requiring demographic balance conditions to be satisfied will distort the market for data and eventually create utility losses for all parties involved (i.e., data consumers, data producers and the data exchange platform). If that loss remains large, no data exchange would risk operating in a non-competitive way and regulation, while effective in view of fairness, may simply be harder to argue for and deploy. Alternatively, if we can show under some conditions that this loss is relatively small, there is a stronger argument to intervene in the market to ensure demographically-balanced datasets are produced. Determining which of those cases applies is the motivation behind this paper.

We investigate the effects of intervening for demographic balance in a specialized variant of a data market proposed in [26]. This specialized variant (see Section (3.3) for details) brings training data into the scope of the fairness–efficiency trade-off by endogenizing data production. In the model sellers decide how many training samples to produce for each group, then submit their training data to a centralized marketplace; buyers want to buy predictions, and to do so submit their prediction tasks and a bid that signals their willingness-to-pay to the marketplace; the marketplace produces predictions, sets prices for the buyers, and divides the revenues among the sellers. Each

seller aims to maximize the difference of its share of the revenues and its production costs. Groups may systematically differ along three dimensions that bear on these quantities and may drive the sellers to unfair data production in the absence of a fairness intervention: 1) the potential economic value that can be extracted; 2) the difficulty of the prediction tasks; and 3) marginal training-sample production costs.

We illustrate the model with the following running example. The sellers could be data brokers, firms that collect troves of consumer data and sell that data and derived data products primarily to other firms such as Axciom and Corelogic [70], who enter the market to monetize their data via an additional revenue channel. Data brokers decide how much of their data trove to submit to the marketplace, and whether to collect more data based on market demand. The buyers could be firms of any size and in any industry, as long as they can define consumer prediction tasks that are valuable to them. Firms transact with the marketplace to buy predictions. For example, a firm could be a small specialty custom cane maker whose customers might skew towards older individuals with an interest in martial arts.[1] Although data brokers typically offer consumer-segment data products that may already be valuable to the firm such as "senior products buyer" and "affluent baby boomer" [70], there may be further value to the firm in defining novel consumer prediction tasks based on its own data such as its historical sales across its various product lines, e.g. based on the characteristics of past customers, would a prospective customer be more interested in a cane or a walking stick? The cane maker might be able to estimate the marginal value of accuracy on the prediction task, e.g., it expects a 1% increase in accuracy in correctly predicting a customer's preference between a cane or walking stick to yield $1,000 in additional sales annually. In the absence of a fairness intervention, the participation of the cane maker in the data market may skew data production towards training samples of seniors. If this skew is unacceptable, then a fairness intervention may be needed to ensure that non-seniors are equitably represented. But this raises many questions that we investigate in the model.

An intervention changes the demographics of the training data by design, but also unavoidably

[1] https://canemasters.com/

constrains the extraction of economic value. If fewer revenues are extracted, then the budget for data production decreases and sellers will produce fewer training samples. In contrast to works that assume a fixed budget for data production as in [69, 71], the cost of fairness now arises out of a complex dynamic that includes this feedback and it is unclear what the overall impact will be. Does the feedback amplify the cost of fairness? And if so by how much? Moreover, the cost of fairness is no longer a single quantity borne by a single practitioner, it affects every agent in the data market, every buyer, seller, and the marketplace. How is the cost of fairness distributed among them? And how do their distributions interact? Summarized in one broad question, we ask: *how does the cost of fairness behave in this data market?*

Our results suggest that any analysis of the cost of fairness must take into consideration the economics of the data production and value extraction in the market, leading to its full potential economic value. In a nutshell, our results highlight that the group with the largest potential economic value determines entirely whether fairness intervention succeeds:

- When the potential economic value is small for all groups, the cost of fairness can be unbearable and the market may not form at all. This can harm even the very groups the intervention is intended to benefit: leading to decreased accuracy and data produced for them.

- When alternatively the potential economic value *of at least one of the group* grows unbounded, the intervention can affect agents only in two positive ways. First, some agents can be strictly better off; the intervention can have a positive externality that benefits some market participants in addition to creating more data and higher accuracy for the intended groups. But more surprisingly, for *all* the other market participants, including all data producers, data consumers and the platform, the cost of fairness seen as a fraction of their utilities without intervention decreases, even vanishes and becomes arbitrarily small.

Data markets are often praised for their superior ability to extract value from data but criticized for the concerns they raise around privacy and fairness. While reconciling privacy with efficient data-value extraction has received plenty of attention [28, 72], so far fairness has rarely been considered

except as an obstacle [73, 69]. Our results suggest that value extraction and fairness may not necessarily be at odds: Model-based data markets can efficiently align the efficiency of data-value extraction with a higher mandate to ensure fairness conditions and convert economic growth into opportunities to intervene for fairness. We believe that our results are only the first in that direction, while other market mechanisms and other forms of fairness guarantees could be leveraging interventions that take into account the market's endogenous response.

In the remainder of the introduction, we summarize the rest of the paper to present the contributions sketched above in more details.

After presenting related work (Section 3.2) and our model definition (Section 3.3), we study we study the data-market equilibria under a fixed number of buyers (Section 3.4). We first show that the equilibria cannot be described in closed form in the full generality of the model. Interestingly, this arises from variation in the characteristics of only one group, independent of variation in characteristics between groups. In particular, when buyers can have different prediction tasks associated to the same group. Thus analyzing the cost of fairness in a data market requires attending to intra-group variation as well as inter-group differences, in general. For example, the small cane maker may not be the only firm with an interest in seniors, another firm might be a large fortune 500 pharmacy chain that carries a portfolio of senior-oriented health products including canes. Although many of both firms' customers may be seniors, those senior customers may otherwise be significantly different. To continue studying the data-market equilibria, we narrow our focus to a more restricted *quasi-symmetric* setting of the model in which the buyers face a common prediction task across all groups and the sellers face the same production costs.

In Section 3.4.1 we characterize the equilibria when no fairness intervention is undertaken, i.e., in a *baseline scenario*. At equilibrium, the number of training samples produced for each group is independent of the other groups and based on the group's potential economic value, production costs, and the difficulty of the common prediction task. If the sellers' production costs are too high, then no seller will produce any data for the group, and otherwise, more samples are produced as the group's potential economic value increases and production costs decrease. When samples

are produced for a group, the difficulty of the prediction task exerts a dampening effect with fewer samples being produced as the prediction task becomes easier. This shows how complicated the market dynamics that shape dataset-demographics at equilibrium can be, involving interactions between revenues, costs, and machine learning. And it confirms that ensuring demographically-balanced data production at equilibrium requires active intervention.

In Section 3.4.2 we characterize the equilibria when the marketplace undertakes a simple intervention, i.e., in an *intervention scenario*. The marketplace stipulates the dataset demographics that satisfy demographic balance, and accepts a seller's dataset if and only if the dataset is demographically balanced. This simple intervention subsidizes data production for some groups by coupling the extraction of economic value across all the groups and impacts the total number of training samples produced, the sellers' production costs, and whether sellers can profitably produce training samples. The exact impact on each of these quantities depends on the marketplace's stipulated demographic balance, some marketplace may aim for different values of demographic balance, more or less geared towards equal representation. This indicates that a fairness intervention in a data market may have some flexibility to trade-off fairness against efficiency. The extent of that flexibility, and how it affects market conditions are issues we study in subsequent sections.

First, in Section 3.5 we apply the analyses of the data-market equilibria for a fixed number of buyers in the baseline and intervention scenarios to study how choosing different values of demographic balance affects the cost of fairness. We find that the choice of demographic balance must be made carefully. Specifically, we prove that for every target demographic balance, there exist market conditions under which all of the sellers opt out of the market and produce no training samples at equilibrium when this intervention is enforced. In that case, the cost of fairness is maximum—every agent's entire utility in the baseline scenario is lost under such fairness intervention. Moreover, this can harm the very groups the intervention is intended to benefit as it loses the original utility (however unfair it was initially in comparison to other groups) that was initially produced before fairness intervention.

This immediately motivates us to find in Section 3.6 conditions under which demographic

balance target do not prevent sellers to participate and utility to grow. Ideally one would like to make the cost of fairness as small as possible for a vast range of choices of demographic balance. We find that the risk that sellers will opt out of the market can be mitigated *provided that at least one group has a sufficiently large potential economic value*. More precisely, we turn to investigate the asymptotic impact of an unbounded number of buyers entering the data market, *i.e.* more and more firms being able to extract value from the data, although perhaps with a diminishing return on the value they extract. Surprisingly, we find that if the economic potential of at least one group is asymptotically unbounded, then for every buyers/sellers, either the cost of fairness becomes a vanishing fraction of its baseline utility or it is asymptotically strictly better off. In other words, once a market expands sufficiently, either fairness has no cost, or it brings a positive externality. This result holds asymptotically for any choice of demographic balance, proving that economic potential confers flexibility to the marketplace to intervene for fairness, reducing the burden of fairness for every agent involved.

Our results are based on some strong assumptions and come with some important limitations. On a policy level, we assume that there is a target demographic that suitably captures fairness. This is feasible in some contexts [42], and challenging in others [71]. In our context, it is unclear when a centralized marketplace is well-qualified to determine a suitable target demographic, and if it is not, then what entity should. More fundamentally, we assume that demographic balance is an appropriate fairness criterion which can only be determined in practice by taking relevant factors into account including ethical, societal, and technical, and it is unclear the extent to which our results generalize to other fairness criteria such as demographic parity [37].

On a technical level, we emphasize that our work is only a promising first theoretical step. Here are a few promising directions for future work that our results motivate. The fairness intervention we introduce is simple and may seem naive or unsuitable for direct implementation. Its main merit is to highlight the opportunities that fairness interventions offer in general, and we hope that more subtle fairness intervention improves upon our results. We assume that training samples from each group are completely uninformative for prediction tasks of any other group which may amplify

the economic disparities and unfairness between groups. We also proved that our results hold in the opposite case where group are equally informative to each other but unfairness may rise due to different production costs among groups, but we omit those not to overload the notation. Studying a general model of partial transfer therefore appears promising. And, we model the data market as a simultaneous game which does not take into account the problem the agents face to learn their equilibrium strategies or the sequential nature of market transactions in the real world. Notwithstanding these limitations, overall, we believe that our work provides an exciting novel perspective that highlight further opportunities for intervening to make data production more equitable.

3.2 Related Literature

We primarily study the cost of fairness, i.e., the cost to system performance that is incurred to satisfy a fairness criterion. The cost of fairness is fundamental in fair machine learning and algorithmic fairness; it provides motivation and shapes the objectives of several research genres.

A foundational genre focuses on characterizing the minimum cost of fairness with the aim of supporting methods to achieve that minimum. Menon and Williamson [37] study the cost of fairness in binary classification. They show that one difficulty in minimizing the cost of fairness is that it is problem instance-dependent. Corbett-Davies and Goel [74] relate the cost of fairness of several classes of fairness criteria to infra-marginal statistical properties; they show that the minimum cost of fairness can be unacceptably high, even when instance-optimal classifiers are deployed. Dwork et al. [35] study the cost of fairness that arises from the interaction of satisfying an individual fairness criterion and a group fairness criterion simultaneously. Thus showing that the challenges associated with the minimum cost of fairness can intensify when the number of fairness criteria are scaled.

Like these works, we are primarily concerned with the cost of fairness. However, our focus is not on *minimizing* the cost of fairness, but rather on *amortizing* the cost of fairness. Our aim is to exhibit and elucidate for the first time the *amortization of the cost of fairness*, the phenomenon in

which the cost of fairness is amortized to a vanishing fraction over the long-term.

A complementary genre focuses on developing strategies to mitigate the cost of fairness. One strategy is to propose relaxations of existing fairness criteria or fairness criteria with a policy parameter that controls the trade-off between fairness and performance as in [50, 42, 71, 40]. Another strategy is to propose an alternative fairness criterion in place of an original one as in [67, 65, 66, 64, 68]. The alternative gives up on the fairness guarantee of the original, in exchange it provides a qualitatively different guarantee and obtains increased system performance. This strategy is often applied in settings where the original fairness criterion may be inappropriate, and the alternative may provide a more practical option.

Like these works we aim to mitigate the cost of fairness. However, we deploy a novel strategy. Rather than relax or propose a fairness criterion to lower the minimum cost of fairness, we take an existing fairness criterion and *without modifying it*, mitigate the cost of fairness by amortizing it over the long term.

Even when options are available to lower the cost of fairness, there is still an incentive for at least one agent to transgress fairness. A broader, policy-oriented genre encompasses the first two and focuses on the perspective of an external policy-maker. Empirical works in this genre often audit real-world systems to exhibit and study any ostensive unfairness they uncover, as in [75, 43, 76]. Theoretical works in this genre often focus on understanding the need and impacts of external intervention by policy-makers. Liu et al. [77] show that policy-makers must consider the long-term effects of a proposed fairness intervention Elzayn and Fish [69] show that policy-makers need to monitor data markets for machine learning for unfairness and potential intervention.

Our work contributes to the theoretical stream, we show that policy makers should also pay attention to the economic characteristics of the data market to understand the impact and severity of a fairness intervention.

3.3 Model

We revisit a data market proposed in Agarwal et al. [26] and formulate a specialized variant.

3.3.1 Centralized market structure

Our model has a centralized structure in which buyers and sellers interact directly only with a central marketplace. We design a data market with a centralized market structure because centralized market structures occur across a wide range of real-world data markets and offer a number of features that are suitable for considering and studying a fairness intervention. Our model exploits three of these features.

First, centralized market structures can lower transaction costs. Many real-world online data markets lower the search costs of dataset buyers and sellers by functioning as centralized dataset storefronts [22]. Sellers register their dataset with the centralized marketplace; the marketplace implements search functionality over the datasets; buyers use the search functionality to look for suitable datasets. Prominent examples include AWS Data Exchange, Snowflake Marketplace, and Dawex Data Marketplace [78, 79, 80].

A second advantage is that centralized market structures allow the marketplace to refine the sellers' data into more sophisticated data products for the buyers. Generative AI provides a rich class of real-world examples including Amazon's Bedrock, Google's Gemini, Stability AI, and Jammable [81, 82, 83, 84]. We illustrate with the example of OpenAI's chatbot ChatGPT [85]. Sellers sell natural-language data to OpenAI. OpenAI refines the sellers' data into a large language model that drives ChatGPT's responses.[2] Buyers prompt ChatGPT and buy its responses; in machine-learning terms, the buyers submit prediction tasks to ChatGPT and buy its predictions. Such data markets can generate economic surplus to the buyers by making refined data products available that would otherwise be out of their reach. And such data markets can generate economic surplus to the sellers and the marketplace by monetizing the data product and spreading the cost across many buyers.

The third advantage shows that the benefits can extend beyond efficiency. A centralized market structure provides opportunities for regulators to intervene. A regulator can implement a control

[2] As of this writing, ChatGPT is more than a chatbot with various versions that include multimodal capabilities. For simplicity we focus only on its text capabilities, in terms of our discussion of data market structure, the principles are the same.

itself. Although perhaps not the first example to come to mind as a data market, the United States Census Bureau coordinates the production and distribution of data on the people and economy of the United States [86]. The Census Bureau collects vast amounts of survey data directly from individuals, processes the survey data into data products, and makes those data products freely and publicly available [87]. The Census Bureau is also legally obligated to maintain the confidentiality of individuals' survey responses [88]. In addition to policy controls, the Census Bureau uses statistical techniques to protect the privacy of individuals while preserving the usefulness and accuracy of aggregate statistics [89].

Alternatively, a regulator can implement controls indirectly by enacting legal mandates. The Fair Credit Reporting Act (FCRA) illustrates this approach. The FCRA regulates data markets in consumer reports. The market structure is organized around individual credit reporting agencies, such as Equifax, Transunion, and ADP. A credit reporting agency collects consumer information from multiple sources including creditors, collection agencies, and employers. The credit reporting agency then processes the consumer information into consumer reports and sells those consumer reports. The FCRA imposes legal obligations on the sources of consumer information, the credit reporting agencies, and the users of consumer reports. For example, the credit reporting agencies can sell consumer reports only for purposes that are specifically allowed by the FCRA. The FCRA shows that the central marketplace can be effectively regulated even in the presence of competition, e.g., Equifax and Experian are competitors.

Overall, centralized market structures can enable both efficiency and regulation. Therefore, they make a sensible starting point for studying the economics of fairness interventions in data markets. Our model reflects a centralized market structure as follows. There are three kinds of agents: 1) M sellers that produce and sell data to the marketplace; 2) N buyers who buy predictions from the marketplace; and 3) a marketplace that aggregates and allocates data, produces predictions, and sets prices.

3.3.2 Data supply and the sellers

A key innovation of our model is that it endogenizes data production. Sellers respond to market conditions by deciding whether to produce data and, if so, what data to produce. In this section we develop our models of data-supply and the sellers.

Data-supply Model We model data supply—the form in which sellers bring data to market—to address a number of issues.

First, data supply determines what data products are possible. For example, if data supply is restricted to a list of sensor observations, then the model can only capture data markets that trade in sensor observations or data products derived from sensor observations [23].[3]

Second, data supply shapes the extraction of economic value. In order to complete a market transaction, the data supply has to be refined into a data product, and the final data product has to be priced.

Third, data supply shapes the notions of fairness that can be applied in the model. We study a notion of fairness that is based on the data supply, we defer the main discussion of our fairness model to Section (3.3.6). In addressing these issues, our goal is to model data supply to capture as large and diverse a universe of data markets as possible, both real-world and theoretical.

Data supply in data markets can vary greatly along many dimensions. To give a few examples, one can find demographic data in a tabular, CSV format from the Census Bureau, job-posting data in a non-tabular, JSONL, format on AWS Data Exchange, and speech data in audio clips, MP3s, on Mozilla Common Voice [90, 91, 92]. These illustrate how data supply can vary in application domain, content, and digital format.

We model data supply to allow for this variation by focusing on a common denominator: there is usually a fundamental notion of a *sample*. A sample organizes pieces of information in the dataset into a discrete unit. For example, a row in a CSV file is the data related to a particular individual, an object in a JSONL file is the data related to a job post, or one MP3 file in a directory

[3]That is perfect when the goal is to design a mobile crowd-sensing data market as in [23]; our goal is to study fairness interventions in data markets as broadly as possible.

of MP3 files is one utterance of a sentence. By focusing on the sample level, we can abstract away such lower-level details that unnecessarily complicate our analyses or limit the scope of our model.

The notion of a sample gives rise to the natural notion of *dataset size*, i.e., how many samples there are in a dataset. Dataset size provides a natural dimension to compare datasets that may be incomparable along other dimensions such as application domain, content, and format. Moreover, the sample level also allows us to retain selected lower-level information.

In order to model fairness, we model samples as belonging to some group. The main discussion of our fairness model is in Section (3.3.6). Here we note that the notion of a group does not restrict the model to data markets that trade in data on people; it is often used to capture demographic groups, particularly ones with legal protections such as age, race, and gender [59, 67], but it is also applied more broadly to non-demographic groups such as languages and countries [71, 93].

Putting this all together, we model data supply as datasets. A dataset is a number of samples where each sample is associated to some group. Besides being supported by the rationale we have just discussed, this is a common model in the literature on fair machine learning [69, 71]. Moreover, dataset size is intimately connected to generalization error empirically and theoretically under the assumption that the samples are independent and identically distributed [7, 8, 9]; we discuss this connection in more detail when we develop our model of data demand in Section (3.3.3). We capture datasets mathematically as follows.

Definition 3.3.1. *(Dataset) A dataset is made up of samples; each sample is exclusively associated to one group g, from a possible set of groups G. The dataset is described by a vector $x \in \mathbb{R}^{|G|}$ where $x_g \geq 0$ gives the number of samples associated to group g in the dataset x. We denote the total number of samples in x, by $\|x\| \triangleq \sum_{g \in G} x_g$.*

Seller Model Sellers produce datasets, each seller decides for itself how many samples to produce of each group. This includes the possibility that it decides to produce no data, i.e., zero samples for every group. We model three factors that weigh on a seller's production decision: accuracy, competition, and production costs. We model accuracy and competition as based primarily

on the buyers and the marketplace, respectively. We model production, however, as a characteristic of the sellers.[4]

Production in real-world data markets is comprised of many factors. An obvious and critical factor is infrastructure; Mozilla's Common Voice Project states that "[it] costs almost a million dollars a year to host the datasets and improve the platform for the 100+ language communities who rely on what we do" [94]. The US Census Bureau's Fiscal Year 2025 Budget Summary lists less obvious but equally critical factors such as quality assurance, testing, and contracting [95]. In the business sector, LiveRamp mentions compliance and licensing factors in its 2024 Annual Report [96]. These examples illustrate that production factors can vary tremendously across data markets.

We aim to parsimoniously model production to capture as much of this variation as possible. Given that the sellers' main goal is to maximize their profits, what ultimately matters across all the variation in production factors is the final monetary production cost of a dataset. We can achieve our aim by focusing on the relationship between a dataset and its monetary cost, i.e., by modeling production costs as a real-valued function of a dataset.

We use a simple model of production costs that is common in the literature on fair machine learning [69, 71]. In this model, the cost of producing a dataset, x, is broken down by group. The cost for producing x_g samples of a particular group $g \in G$ is $\kappa_g x_g$, i.e., the number of samples multiplied by a group-specific constant $\kappa_g > 0$. The cost of producing x is the sum of the group-specific costs.

This production-cost model is based on three assumptions that limit the scope of our data-market model. First, we assume that costs can be attributed to individual samples. Second, we assume that these costs are additive. And third we assume that within each group, every sample has the same production cost.

This model has clear limitations, yet we believe that it makes a sensible first step for our investigation of the cost of fairness in data markets. We leave the study of more sophisticated

[4]We develop our model of accuracy in Section (3.3.3); competition in Section (3.3.4); and tie them together with production in the seller's decision-making in Section (3.3.5).

production-cost models to future work.

Definition 3.3.2. *(Sellers) There are M sellers. Each seller j produces a dataset $x^{(j)}$ to sell. We assume that, for each group g, seller j's production process ensures that the $x_g^{(j)}$ samples in the dataset are independent and identically distributed, and that samples are drawn independently between groups.*

Each seller j faces a production-cost structure, $\kappa^{(j)} \in \mathbb{R}^{|G|}$; $\kappa_g^{(j)} > 0$ is the constant marginal cost incurred by seller j to produce a sample of group g. The cost to seller j of producing dataset $x^{(j)}$ is $\sum \kappa_g^{(j)} x_g^{(j)}$, or in vector notation, $\kappa^{(j)T} x^{(j)}$.

3.3.3 Data demand and the buyers

Buyers There are N buyers. Each buyer i faces $|G|$ prediction tasks, one for each group g. For example, while most of the cane maker's customers may be seniors, many of its customers may also be non-seniors, and the cane maker may be able to obtain improved accuracy by splitting its prediction task between seniors and non-seniors.

The relationship between the number of training samples and accuracy on the prediction task is described by three parameters: $Z_{i,g}$, $\alpha_{i,g}$, and $\beta_{i,g}$. Given n samples from group g, the accuracy, $\mathcal{G}_{i,g}(n)$, of the prediction task is,

$$\mathcal{G}_{i,g}(n) \triangleq \left(Z_{i,g} - \alpha_{i,g} n^{-\beta_{i,g}}\right)_+, \tag{3.1}$$

where $(\cdot)_+$ is the positive part. In addition to its prediction tasks, each buyer i has a values-of-accuracy vector $\mu_i \in \mathbb{R}^{|G|}$ which connects accuracy on the prediction task to economic value to the buyer. Buyer i's willingness-to-pay for accuracy $\mathcal{G}_{i,g}(n)$ is $\mu_{i,g} \cdot \mathcal{G}_{i,g}(n)$ for group g. The buyers' values-of-accuracy are private information and unknown to all the other agents. For example, possibly reflecting the cane maker's core consumers, its value-of-accuracy for seniors might be \$100,000 while its value-of-accuracy for non-seniors might be \$50,000.

We assume that training samples of one group are completely uninformative for the prediction

tasks of any other group.

Assumption 3.3.1. *(Zero inter-group transfer) Let x be a dataset. Then for every buyer i and group g we have,*

$$\mathcal{G}_{i,g}(x) = \mathcal{G}_{i,g}(x_g). \tag{3.2}$$

Note that we are abusing notation here, x is a vector whereas x_g is a scalar, but we hope that this clarifies that when the whole dataset is passed, only the training samples of the group g are informative.

Assumption (3.3.1) is a simplification of the real-world observation that lessons don't necessarily fully transfer across populations. This observation underlies many works in the fairness literature [74, 97]. Indeed, in some settings, if you could learn from just the majority population and transfer to the minority population, fairness would be trivial [74, 36]. Still, in reality, there is often at least some transfer. However, it is suitable for the purposes of this first theoretical work for two reasons: 1) modeling transfer is an interesting issue that raises many questions for future work; and 2) this decouples the extraction of value by groups when there is no fairness intervention and likely amplifies the cost of fairness, this ensures that our striking result that the cost of fairness can vanish is not because fairness turns out to be optimal in some counter-intuitive way.

3.3.4 Market mechanics and the marketplace

We have specified the sellers and buyers. We now turn to the marketplace. The marketplace is best understood in terms of how it coordinates the market. We first explain the role of the marketplace and how it fulfills that role at a high level before specifying the mechanics of the market.

In principle, buyers and sellers could transact directly and do not require a third party. But pricing in bilateral data market transactions is challenging. The buyers and sellers face Arrow's paradox: a buyer cannot know what it is willing to pay without seeing the data, but once it has seen the data it has already consumed the good and has no incentive to pay the seller, so a seller will not

disclose the data to the buyer before they agree to a price [98]. A data market must solve Arrow's paradox in some way.

One common solution in real-world online data markets is the use of posted prices [24, 22]. Data sellers post a price on the online platform that buyers can accept or reject. The widespread use of posted prices indicates that this is a broadly viable solution. But it does not always work perfectly. Some real-world online data markets have ostensibly posted prices; although sellers post a price, once buyers find a seller, they often continue the transaction bilaterally and offline—including price negotations [22].

Our model, based on that of Agarwal et al. [26], envisions a different solution to Arrow's paradox, one that leverages the central position of an online platform and is based on two key ideas: 1) the marketplace sells predictions, rather than data, to the buyers; and 2) transactions are fully intermediated by the marketplace, the buyers and sellers never interact directly.

In essence, the sellers and buyers delegate the tasks of machine learning and pricing to the marketplace. The sellers give the marketplace access to their datasets and the buyers disclose to the marketplace their values-of-accuracy and prediction tasks. The role of the marketplace is to perform three functions: 1) produce predictions; 2) set prices; and 3) divide payment between the sellers. Thus, a price that reflects the value of the data can be set without the buyers learning about the data before paying. Although this solves Arrow's paradox, it raises another challenge for the marketplace.

Given access to the sellers' datasets, the marketplace can straightforwardly implement this solution *if it accurately knows* a buyer's values-of-accuracy and prediction tasks. But this information is private to each buyer. A buyer may not accurately disclose its values-of-accuracy and prediction tasks if it anticipates that doing so will favorably affect the price [99]. In other words, the marketplace faces the problem of truthfully eliciting the buyers' private information.

In their blueprint for an online data market, Agarwal et al. [26] solve this problem by using the celebrated Myerson's mechanism [100]. Myseron's mechanism was originally developed in the context of auctions for the purpose of eliciting truthful bids. The mechanism consists of two

components: an allocation function and a revenue function. On input the bidders' bids, the allocation function specifies how much good each bidder receives. The revenue function is constructed from the allocation function such that truthful bidding is a dominant strategy for every bidder. A key innovation of Agarwal et al. [26] is to structure the interaction between the marketplace and each buyer so that the marketplace can apply Myerson's mechanism. We explain the full buyer–marketplace interaction in detail below, here we focus on the difference between our model and that of Agarwal et al. [26].

From the perspective of modeling a data market, the designable component of Myerson's mechanism is the allocation function. The mechanism can be implemented using any monotonic allocation function, i.e., a bidder never gets less for bidding more. Agarwal et al. [26] exploit much of this flexibility in their model, they focus on a class of reserve-price based allocation functions. These allocation functions take a reserve price in addition to a buyer's bid. If the buyer bids at least the reserve price, the full dataset is allocated for machine learning. Otherwise, the dataset is degraded in some way, e.g., less than all the samples are allocated or noise is added to the dataset.

In our model, we restrict the marketplace to an all-or-nothing allocation function: for each group, the buyer gets all the data if it bids at least the reserve price for that group and nothing otherwise. This restriction serves several purposes. First, it is a common benchmark in prior-free settings where an auctioneer has no distributional information about buyer valuations [28].

Second, it enables insightful theoretical analyses. As we will show in Section (3.4), even this restriction is analytically intractable in a moderately complex regime. We will then make a further restriction to continue our analyses. Although there are other ways one could restrict the allocation function for theoretical analyses, we believe this restriction is an appropriate one for the next reason.

Our restriction bridges the gap between real-world online data markets with posted prices and the blueprint of Agarwal et al. [26]. It generalizes the posted price mechanism and specializes Myerson's mechanism. Thus, our analyses capture: 1) many real-world online data markets as a special case of our model; and 2) a special case of the model of Agarwal et al. [26]. We believe

that striking this balance between real-world and theoretical data markets evidences the robustness of our results on fairness in data production.

Myerson's mechanism incentivizes the buyers to truthfully report their values-of-accuracy. But a buyer must also report its prediction tasks, which independently enter into its willigness-to-pay. We follow Agarwal et al. [26] in assuming that the buyers do not misreport their prediction tasks. It seems much harder and riskier for a buyer to effectively misreport its prediction tasks. Doing so likely requires the buyer to have some information about how the datasets relate to its prediction tasks. But the buyers never have access to the datasets. Thus, the buyers would have to infer this relationship, and this opens up the possibility that strategically misreporting its prediction tasks may backfire—it may end up lowering the quality of the predictions it receives. Still, it is not clear that a buyer may never be better off by strategically misreporting its prediction tasks. We leave this issue to future work.

We briefly summarize the marketplace's role before moving on to the mechanics of the market. The data market must solve Arrow's paradox. The marketplace solves Arrow's paradox by fully intermediating market transactions, the marketplace: 1) executes the machine learning; 2) sets prices; and 3) divides payments among the sellers. This solution, comes with another challenge: the marketplace needs to accurately know the buyers' private information. The marketplace uses Myerson's mechanism to incentivize the buyers to submit truthful reports.

Market Mechanics Market transactions begin with an interaction between the sellers and the marketplace. Each seller j gives the marketplace access to its dataset, $x^{(j)}$. The marketplace can combine the sellers' datasets for machine learning and revenue division. For a subset of sellers $T \subseteq [M]$, the dataset $x^{(T)}$ is defined coordinate-wise by,[5]

$$x_g^{(T)} = \sum_{j \in T} x_g^{(j)}. \tag{3.3}$$

[5]We are abusing notation here; this means that seller j's dataset could also be written as $x^{(\{j\})}$, but for the sake of notational clarity we will only use $x^{(j)}$.

In particular, $x^{([M])}$ is the aggregate dataset of all the sellers' datasets. With access to the sellers' data, the marketplace has one of the key production factors for predictions.

Next, the marketplace chooses a reserve-price vector, $p \in \mathbb{R}^{|G|}$, $p_g > 0$, and then publicly announces p so that buyers know p at the beginning of the next step.

Once p is announced, each buyer i submits its prediction tasks and a bid vector $b_i \in \mathbb{R}^{|G|}$ to the marketplace. The bid vector b_i is buyer i's report to the marketplace of its values-of-accuracy μ_i.

Then, the marketplace allocates training samples from the aggregate dataset to carry out machine learning and produce predictions for the buyers' prediction tasks. The marketplace uses a reserve price allocation mechanism, $\mathcal{AF}_g$, for each group g. For each group g, the marketplace allocates all the samples in the aggregate dataset to the machine learning for buyer i's prediction task for group g if the buyer bids at least the reserve price p_g, and otherwise nothing. Formally,

$$\mathcal{AF}_g(b_{i,g}, x^{([M])}) \triangleq \begin{cases} x_g^{([M])} & \text{if } b_{i,g} \geq p_g \\ 0 & \text{if } b_{i,g} < p_g \end{cases} \tag{3.4}$$

Note that we define $\mathcal{AF}_g(b_{i,g}, x^{([M])})$ to be $x_g^{([M])}$ rather than $x^{([M])}$ when $b_{i,g} \geq p_g$ for ease of presentation due to the zero inter-group transfer assumption. Having allocated training samples for machine learning to buyer i's prediction task for group g, the marketplace then carries out the machine learning to produce predictions and sets a price for the predictions of

$$\mathcal{RF}_{i,g}(b_{i,g}) \triangleq p_g \mathcal{G}(\mathcal{AF}_g(b_{i,g}, x^{([M])})). \tag{3.5}$$

Note that this is the revenue function required by Myerson's mechanism for the allocation function $\mathcal{AF}_g$.

Finally, the marketplace divides the collected revenues between the sellers using the Shapley value. For notational clarity, we define $c_T \triangleq |T|!(M - |T| - 1)!/M!$ for any coalition $T \subseteq [M]$.

For each buyer i and group g, seller j receives the payment division

$$\mathcal{PD}_{i,g,j}(x^{(j)}) = p_g \sum_{T \subseteq [M] \setminus \{j\}} c_T \cdot \left(\mathcal{G}_{i,g}(\mathcal{AF}_g(b_{i,g}, x^{(T \cup \{j\})})) - \mathcal{G}_{i,g}(\mathcal{AF}_g(b_{i,g}, x^{(T)})) \right). \quad (3.6)$$

3.3.5 Market outcomes: equilibria, formation, and growth

We have specified the various market participants and the mechanics of the data market. We want to study the behavior of the market in terms of the outcomes it may generate. How much data will be produced? How much of the data's value will be unlocked? How much revenues will be collected? And how will these outcomes differ when a fairness intervention is implemented?

Our fundamental approach to investigating these questions is to model the data market as a simultaneous game and study the Nash equilibria of the market.

Utilities and Equilibria Agents in the data market are strategic and act to maximize their utilities. The marketplace chooses the reserve-price vector p so as to maximize its total revenues,

$$w(p) \triangleq \sum_{i=1}^{N} \sum_{g \in G} \mathcal{RF}_{i,g}(b_{i,g}). \quad (3.7)$$

Each seller j produces a dataset $x^{(j)}$ to maximize its profits,

$$v_j(x^{(j)}) \triangleq \left(\sum_{i=1}^{N} \sum_{g \in G} \mathcal{PD}_{i,g,j}(x^{(j)}) \right) - \kappa^{(j)T} x^{(j)}. \quad (3.8)$$

Each buyer submits bids b_i to maximize its surplus,

$$u_i(b_i) \triangleq \sum_{g \in G} \mu_{i,g} \mathcal{G}_{i,g}(\mathcal{AF}_g(b_{i,g}, x^{([M])})) - \mathcal{RF}_{i,g}(b_{i,g}) \quad (3.9)$$

$$= \sum_{g \in G} (\mu_{i,g} - p_g) \mathcal{G}_{i,g}(\mathcal{AF}_g(b_{i,g}, x^{([M])})). \quad (3.10)$$

A strategy profile $(p, \{b_i\}, \{x^{(j)}\})$ is a Nash equilibrium if no agent can improve its utility by a unilateral deviation in its strategy.

Market formation We aim to investigate the impact of intervening for fairness on data supply. Towards this end, our model endogenizes data supply. It does not assume that there is a fixed amount of data that is exogenously given, rather the amount of data supplied depends on the sellers' ability to produce profitably. Crucially, this allows for the possibility that *no data* is produced at any equilibrium. When this happens, we say that the data market does not form. And we say that the data market forms if there exists a Nash equilibrium at which some data is produced. We find in Section (3.5) that a fairness intervention can affect market formation.

Market Growth We also aim to investigate the relationship between intervening for fairness and market growth. We model market growth as markets with more buyers. In the quasi-symmetric setting (see Definition (3.4.1)) where we analyze market growth, the buyers will only differ on their values-of-accuracy vectors. For a fixed number of sellers M, therefore, a sequence of values-of-accuracy vectors $(\mu_N)_{N\in\mathbb{N}}$ defines a sequence of buyers and in turn a sequence of data markets. The N-th data market is defined by the M sellers and the first N buyers in the buyers sequence. We examine the Nash equilibrium outcomes of the individual data markets as well as in the limit as an unbounded number of buyers enter the data market. And we analyze the comparative statics of the limit equilibrium outcomes between the baseline and intervention scenarios.

3.3.6 Fairness in the data market

We have developed a data market: There are buyers and sellers who can buy predictions and sell datasets intermediated by a central marketplace. And we have defined the aspects of the data market that we wish to analyze. Nash equlibria, market formation, and market growth. We now develop our model of fairness in the data market.

The fairness model performs three critical functions. First, it provides a criterion that captures the notion of fairness that we study and allows one to test whether fairness has been achieved.

Second, it stipulates a particular intervention that the marketplace can undertake to achieve fairness in the intervention scenario. And, third, it allows the cost of the fairness intervention to be assessed. We develop each in turn below.

Fairness Criterion The choice of fairness criterion is a fundamental issue. Many considerations must be taken into account including ethical, technical, and cost. Fundamentally, the fairness criterion expresses the moral imperative of the underlying ethical considerations and expands fairness from the realm of pure rhetoric into the quantifiable. Technical and cost considerations do not substitute ethical considerations, but they can impact the suitability or feasibility of achieving a fairness criterion in a given system.

The marketplace is in a commanding position, it controls several stages of the machine-learning pipeline including pre-processing, in-processing, and post-processing [38]. From a purely technical perspective, the marketplace has many options, it can choose from a wide range of fairness criteria. Here, we explore just one fairness criterion that is based on the fraction of samples associated to each group in a dataset, i.e., the dataset demographics.

Definition 3.3.3. *(Dataset demographics) The* demographics of a dataset x *is the vector* $\gamma(x) \in \mathbb{R}^{|G|}$ *whose* g*-th coordinate,* $\gamma_g(x)$ *is given by,*

$$\gamma_g(x) \triangleq \frac{x_g}{||x||}. \tag{3.11}$$

It is well-understood that the demographics of a dataset are important for fairness. Sufficiently balanced demographics are important for conducting equity assessments of an algorithmic system [68, 34]. Unbalanced demographics can unfairly favor machine-learning performance on one group over another [34]. Balanced demographics may be an important objective in itself that supports the legitimacy of a decision-making process, as in participatory budgeting [42]. Implicit here is a notion that some demographics are balanced, and therefore may be considered fair, while others are unbalanced, and may be considered unfair. We make this notion formally explicit as follows.

Definition 3.3.4. *(Demographic balance) Let* $\gamma \in [0,1]^{|G|}$ *be a* target vector *satisfying* $\sum_{g \in G} \gamma_g = 1$. *We say that a* dataset x is γ-demographically balanced *if for every* g *it holds that*

$$\gamma_g = \frac{x_g}{\|x\|}. \tag{3.12}$$

Note that we are overloading notation here, we use standalone γ *to refer to a target vector and function invocation* $\gamma(x)$ *to refer to the demographics of the dataset* x.

At the same time, the relationship between dataset demographics and fairness is complicated. It is not always clear which target vector is appropriate—indeed, it may be contested. Although relaxing demographic balance to allow for some subset of target vectors might ease the difficulty, it does not address the fundamental ethical issue: What is a fair dataset demographics? Even if there is no difficulty in selecting a target vector, demographic balance in itself does not guarantee fairness. Indeed, it does not even guarantee improved fairness because the dataset demographics propagate downstream in ways that can be subtle and difficult to anticipate [34]. Considered in isolation, demographic balance raises serious concerns.

From a purely fairness perspective, it would be ideal to allow the use of multiple fairness criteria, one that is appropriate for each buyer. One of the reasons we chose to study this model is because the marketplace is positioned to technically implement this vision, at least in principle. And we believe that this is an exciting direction for future work, but this also raises significant challenges.

First and foremost, this raises the problem of choosing a fairness criterion for every buyer. Who will decide? The buyers, the marketplace, or an external regulatory authority? If the buyers decide, what assurance is there that their decision is actually fair to the downstream stakeholders in their application context? If the marketplace or an external regulatory authority decides, how are they to do so? And why are they the right judge and enforcer of fairness?

These questions also apply to choosing a single fairness criterion to enforce across all the buyers. Doing so assumes that the chosen criterion is appropriate, or at least acceptable, in the

application of every buyer. But buyers in real-world data markets face different fairness problems with different fairness, technical, and legal considerations. These difficulties can only grow as the data market seeks to innovate and scale [96].

Choosing demographic balance does not obviate these difficulties. However, in comparison to most fairness criteria in the literature, demographic balance, is lightweight. It does not require the marketplace to change the outputs of machine learning, i.e., the models and their predictions. Although this foregoes the advantages such criteria enjoy—they take principled positions on fairness and make substantive changes to machine-learning outcomes—it also mitigates the severe disadvantages just discussed.

And demographic balance has a further advantage in this context. We are studying the relationship between data supply and a fairness intervention. Demographic balance necessitates an intervention directly on the data supply. Given all of the above considerations, we believe that demographic balance is a sensible criterion to study as an initial step.

Fairness Intervention Settling on a fairness criterion is an essential first step towards building fair machine-learning systems. The next step is to implement a fairness intervention. To achieve demographic balance in the data market, the marketplace must do three things: 1) choose a target vector γ; 2) choose a dataset to target; and 3) choose how to intervene.

Since our aim is to understand intervening for fairness to achieve demographic balance over all possible target vectors, we treat the target vector γ as an exogenous given.

There are a number of datasets that the marketplace could focus on. Each of the sellers' datasets and the aggregate dataset jump out. Since the marketplace allocates the aggregate dataset to any buyer that bids the reserve price, we suppose that the marketplace primarily targets the aggregate dataset.

Finally, we suppose that the marketplace implements the following intervention. On input, a target vector γ, the marketplace accepts a seller j's dataset $x^{(j)}$ if and only if $x^{(j)}$ is γ-demographically balanced. Essentially, the marketplace conditions seller participation on their datasets being γ-

demographically balanced.

If a seller produces any data at equilibrium, its dataset will be demographicaly balanced. The seller's decision is reduced to the total number of samples it will produce, i.e., $\|x^{(j)}\|$. For notational convenience, we define $n^{(j)} \triangleq \|x^{(j)}\|$ for a single seller j and $n^{(T)} \triangleq \sum_{k \in T} n^{(k)}$ for a coalition of sellers T. In particular, $n^{([M])} = \|x^{([M])}\|$ is the total number of samples produced by all the sellers in the aggregate dataset. Therefore, if any seller produces data, the aggregate dataset will be demographically balanced.

This naive intervention is unsuitable for practical application. It is clearly inefficient, for example, it does not allow for sellers to specialize in producing samples for specific groups. Therefore, it likely over-estimates the cost of fairness (see details below). We leave designing and studying an optimally efficient intervention to future work. Still, this intervention is suitable for our purposes.

Our overarching objective is to study fairness interventions in data markets. One of our supporting goals is to understand the behavior of the cost of fairness and, in particular, the relative cost of fairness as the market grows. Because our naive intervention likely inflates these quantities, we argue that it indicates the robustness of our results on the amortization of the cost of fairness (See Section (3.6)). Our results should qualitatively hold for any intervention that is more efficient including the optimally efficient intervention.

The Cost of Fairness The cost of fairness quantifies the burden imposed by a fairness intervention. It is typically defined as the difference between an agent's utility without fairness requirements and its utility with fairness requirements. In contrast to works in fair machine learning that focus on the utility of the practitioner where the fairness intervention can only potentially leave the practitioner worse off, we find that in a data market some agents can be strictly better off. This, perhaps awkwardly, yields a negative cost of fairness for those agents. Therefore we focus on the ratio of an agent's intervention utility to its baseline utility which expresses this contingency more naturally as a ratio greater than 1. The ratio also captures the burden of a fairness intervention in normalized terms so that different data markets can be compared.

Let $(p, \{b_i\}, \{x^{(j)}\})$ be a Nash equilibrium in the baseline scenario, and $(p^f, \{b_i^f\}, \{y^{(j)}\})$ be a Nash equilibrium in the intervention scenario. The marketplace's utility ratio is,

$$UR_{Mkt}(p, p^f) \triangleq \frac{w^f(p^f)}{w(p)}, \tag{3.13}$$

where $w^f(p^f)$ is the marketplace's utility in the intervention scenario, i.e., Equation (3.7) evaluated with respect to the intervention-scenario Nash equilibrium. Seller j's utility ratio is,

$$UR_{S,j}(x^{(j)}, y^{(j)}) \triangleq \frac{v_j^f(y^{(j)})}{v_j(x^{(j)})}, \tag{3.14}$$

where $v_j^f(y^{(j)})$ is seller j's utility in the intervention scenario, i.e., Equation (3.8) evaluated with respect to the intervention-scenario Nash equilibrium. Buyer i's utility ratio is,

$$UR_{B,i}(b_i, b_i^f) \triangleq \frac{u_i^f(b_i^f)}{u_i(b_i)}, \tag{3.15}$$

where $u_i^f(b_i^f)$ is buyer i's utility in the intervention scenario, i.e., Equation (3.9) evaluated with respect to the intervention-scenario Nash equilibrium.

3.4 Data market equilibria under N buyers

In this section we study the data market equilibria when the number of buyers is fixed. Our first result is that the equilibria in the general setting of the model cannot be described in closed form.

Proposition 3.4.1. *There does not exist a general closed-form solution over all the possible equilibrium equations in the general setting of the model.*

Proof. Consider the seller's marginal utility at equilibrium.

Observation 3.4.1. *Fix the bid b_i of each buyer $i \in [N]$, the marketplace's posted price vector p, and the dataset $x^{(k)}$ of each seller $k \in [M] \setminus \{j\}$. Let $x^{(j)}$ be seller j's best response in the*

baseline scenario. For each group g, if $x_g^{(j)} > 0$, then $x_g^{(j)}$ satisfies the equation,

$$p_g \sum_{i \in B_g} \sum_{T \subseteq [M] \setminus \{j\}} c_T \alpha_{i,g} \beta_{i,g} (x_g^{(T)} + x_g^{(j)})^{-\beta_{i,g}-1} = \kappa_g^{(j)}, \tag{3.16}$$

where $B_g \triangleq \{i \in [N] : b_{i,g} \geq p_g\}$ is the set of buyers that bid at least p_g for group g.

Equation (3.16) indicates that the set of possible equilibrium equations includes a large class of polynomial equations that includes polynomial equations of arbitrarily large degree.

The celebrated Abel–Ruffini Theorem [101] tells us that there is no general closed-form solution to polynomial equations of degree five or higher. We cannot immediately apply the theorem, however, because the coefficients that may appear in the equilibrium polynomial equations are not unrestricted. Do polynomial equations without closed-form solutions occur in the equilibrium equations in the general setting of the model? Yes, as demonstrated by the following toy example.

Example 3.4.1. *(van der Waerden's Quintic) Consider the following instance of the model. Let there be 2 buyers and 1 seller. For some group g, set the parameters as follows: $\mu_{1,g} = 1 = \mu_{2,g}$, $Z_{1,g} = 5 = Z_{2,g}$, $\alpha_{1,g} = 1/3$, $\beta_{1,g} = 3$, $\alpha_{2,g} = 1/4$, $\beta_{2,g} = 4$, and $\kappa_g^{(1)} = 1$.*

We know that in the baseline, the marketplace and seller will choose their strategies independently for each group. Since there are 2 buyers each with the same value-of-accuracy for this group, at equilibrium, the marketplace will set $p_g = \mu_{1,g} = \mu_{2,g} = 1$. And since there is only one seller, Equation (3.16) becomes

$$\sum_{i \in \{1,2\}} \alpha_{i,g} \beta_{i,g} (x_g^{(1)})^{-\beta_{i,g}-1} = \kappa_g^{(1)}. \tag{3.17}$$

Multiplying both sides by $(x_g^{(1)})^5$ and rearranging yields the following polynomial equation:

$$(x_g^{(1)})^5 - (x_g^{(1)}) - 1 = 0, \tag{3.18}$$

which is an example given in [102] as having no closed-form solution. Still, the seller will only

solve Equation (3.18) exactly at equilibrium if it can achieve positive utility. The root of Equation (3.18) is approximately 1.1673*, thus we can bound the utility the seller will receive from group g by the prediction gain of producing 1 sample and the cost of producing 2 samples*

$$\sum_{i\in\{1,2\}} p_g \mathcal{G}(x_g^{(1)}) - \kappa_g^{(1)} x_g^{(1)} \geq \sum_{i\in\{1,2\}} \mathcal{G}_{i,g}(1) - 2 = \left(5 - \frac{1}{3}\right) + \left(5 - \frac{1}{4}\right) - 2 > 0. \tag{3.19}$$

We conclude that at equilibrium $x_g^{(1)} > 0$, hence the seller solves Equation (3.18).

□

We see that the impossibility arises in the general setting because buyers can have different prediction tasks *for the same group*. The intra-group variation complicates the possible equilibrium equations that the sellers may face and gives rise to equilibrium equations for which it is known that there is no closed-form solution. This clarifies that the difficulty is independent of inter-group differences, and therefore, to understand fairness in data markets with multiple prediction tasks, it may be important to take intra-group variation into account. But this is not the focus of the present work, we leave this important issue to future work. Despite the impossibility of describing the model equilibria in closed form in the general setting, the more restricted quasi-symmetric setting admits clean and insightful analyses.

Definition 3.4.1. *(Quasi-symmetric setting) The buyers share a common prediction task within groups and between groups, denoted $\mathcal{G}(\cdot)$, and described by parameters Z, α, and β, i.e. for all $i \in [N], g \in G, n \in \mathbb{R}$, $\mathcal{G}_{i,g}(n) = \mathcal{G}(n)$. Although we impose symmetry on the buyers' prediction tasks, we still allow the buyers' values-of-accuracy to vary independently, both between groups and within groups.*

The sellers share a common cost structure denoted κ, i.e., for every pair of sellers $j, j' \in [M]$, $\kappa^{(j)} = \kappa = \kappa^{(j')}$.

The quasi-symmetric setting avoids the difficulties that arise in the general setting by restricting all of the buyers to have the same prediction task within and between groups. We restrict the sellers

to have a common cost structure for ease of presentation, our results qualitatively hold when the cost structures may differ. In the rest of this chapter, we restrict our study and analyses to the quasi-symmetric setting.

We will see that market equilibrium outcomes differ between the baseline and intervention scenarios in general. But the buyers' and the marketplace's equilibrium strategies can be described independently of the scenario. We now describe their strategies before analyzing the market equilibria in each scenario individually.

Buyer Dominant Strategy. The marketplace's allocation function and revenue function are special cases of the model of Agarwal et al. [26], which in turn are a direct application of Myerson's payment function [100]. On this basis, it follows that truthfulness is a dominant strategy for the buyers, i.e., every buyer i always maximizes its utility by bidding $b_i = \mu_i$.

Fact 3.4.1. *(Buyer Truthfulness) For every buyer i, truthfully bidding its values-of-accuracy, i.e., $b_i = \mu_i$ is a dominant strategy.*

Our aim is to study equilibria in the data market. Although truthful bidding is a dominant strategy, there do exist Nash equilibria at which some buyers bid untruthfully. Here is a trivial example, the strategy profile $(0, \{b_i\}, \{0^{(j)}\})$ is a Nash equilibrium for all possible bids $\{b_i\}$, both truthful and untruthful. There exist non-trivial examples as well, but a buyer's ability to bid untruthfully is out of its control, depending on the other buyers and the marketplace, and none of these Nash equilibria improves the utility of an untruthful buyer. Therefore, in the rest of this chapter, we only study strategy profiles in which all the buyers bid truthfully, i.e., strategy profiles of the form $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$.

Marketplace Best Response. Unlike the buyers, the marketplace does not have a dominant strategy. But it does have a best response that depends only on the buyers' strategies and is independent of the sellers' strategies and the scenario.

Let $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ be a strategy profile, and consider the marketplace's utility,

$$w(p) = \sum_{i=1}^{N} \sum_{g \in G} \mathcal{RF}_{i,g}(\mu_{i,g}) \tag{3.20}$$

$$= \sum_{i=1}^{N} \sum_{g \in G} p_g \mathcal{G}(\mathcal{AF}_g(\mu_{i,g}, x^{([M])})) \tag{3.21}$$

$$= \sum_{i=1}^{N} \sum_{g \in G} p_g \mathbf{1}[\mu_{i,g} \geq p_g] \mathcal{G}(x_g^{([M])}) \tag{3.22}$$

$$= \sum_{g \in G} \left(p_g \sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \right) \mathcal{G}(x_g^{([M])}). \tag{3.23}$$

Equation (3.23) shows that the marketplace extracts revenues from each group independently of the others. The revenues extracted from each group g is the product of two factors: one factor is the accuracy $\mathcal{G}(x^{([M])})$; and the other factor will turn out to be critical in our analyses.

Definition 3.4.2. *(Market value-of-accuracy) Let $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ be a strategy profile and g be any group in G. The* market value-of-accuracy of group g in the strategy profile σ, *denoted ρ_g, is defined to be*

$$\rho_g \triangleq p_g \left(\sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \right). \tag{3.24}$$

The market value-of-accuracy, ρ_g is the product of the reserve price and the number of buyers who bid at least the reserve price. ρ_g is the marketplace's marginal revenues in increasing accuracy, in the same way that the buyer value-of-accuracy, $\mu_{i,g}$, is the buyer's marginal economic value in increasing accuracy. Unlike $\mu_{i,g}$, ρ_g is not a fixed and exogenous characteristic of the marketplace, it depends on the marketplace's and the buyers' strategies. However, once the buyers' strategies are fixed, the marketplace effectively controls ρ_g through p_g, and then ρ_g captures how effectively the marketplace extracts marginal revenues. Increasing ρ_g will increase the marketplace's revenues, therefore, the marketplace's best response is to set p_g to maximize ρ_g.

Fact 3.4.2. *(Marketplace Best Response) Let the buyers bid their values-of-accuracy, i.e., $b_i = \mu_i$ for all $i \in [N]$. Let the sellers' datasets $\{x^{(j)}\}$, $j \in [M]$, be arbitrary. Then, the marketplace's*

best response is to maximize ρ_g for every group $g \in G$, i.e., to set reserve prices p_g to,

$$p_g \in \arg\max_p \rho_g. \tag{3.25}$$

Unlike the buyers and the marketplace, the sellers have neither a dominant strategy nor a best response that is independent of the baseline and intervention scenarios. Rather, the sellers' best response differs between the two scenarios because its strategy space differs between the two scenarios; the marketplace restricts the set of datasets the sellers can monetize in the intervention scenario. Therefore, to continue our analyses of the market equilibria, we must study each scenario separately. We next study the baseline scenario.

3.4.1 Baseline Equilibrium

We analyze the sellers' equilibrium strategies. Our first observation is that the sellers produce the same datasets at equilibrium because they face a common cost structure.

Fact 3.4.3. *If $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ is a Nash equilibrium, then for every $j, j' \in [M]$ we have that $x^{(j)} = x^{(j')}$.*

Proof. Seller j's utility is given by

$$v_j(x^{(j)}) \triangleq \left(\sum_{i=1}^{N} \sum_{g \in G} \mathcal{PD}_{i,g,j}(x^{(j)})\right) - \kappa^{(j)T} x^{(j)} \tag{3.26}$$

$$= \left(\sum_{i=1}^{N} \sum_{g \in G} p_g \sum_{T \subseteq [M] \setminus \{j\}} c_T \left(\mathcal{G}(\mathcal{AF}_g(\mu_{i,g}, x^{(T \cup \{j\})})) - \mathcal{G}(\mathcal{AF}_g(\mu_{i,g}, x^{(T)}))\right)\right) - \left(\sum_{g \in G} \kappa_g x_g^{(j)}\right) \tag{3.27}$$

$$= \sum_{g \in G} p_g \sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \sum_{T \subseteq [M] \setminus \{j\}} c_T \left(\mathcal{G}(x_g^{(T \cup \{j\})}) - \mathcal{G}(x_g^{(T)})\right) - \sum_{g \in G} \kappa_g x_g^{(j)} \tag{3.28}$$

$$= \sum_{g \in G} \rho_g \sum_{T \subseteq [M] \setminus \{j\}} c_T \left(\mathcal{G}(x_g^{(T \cup \{j\})}) - \mathcal{G}(x_g^{(T)})\right) - \sum_{g \in G} \kappa_g x_g^{(j)}, \tag{3.29}$$

where $\mathbf{1}[\cdot]$ is the indicator function. At equilibrium, seller j's marginal utility in $x_g^{(j)}$ must be 0 for

every group. Moreover, this must also be true for any other seller k. Hence, at equilibrium, seller j's marginal utility with respect to $x_g^{(j)}$ must equal seller k's marginal utility with respect to $x_g^{(k)}$ for any two sellers j and k.

We now show that if two sellers, j and k produce different amounts of data for any group g, $x_g^{(j)} \neq x_g^{(k)}$, then $(p, \{\mu_i\}, \{x^j\})$ is not a Nash equilibrium. By definition, seller j's marginal utility with respect to $x_g^{(j)}$ is

$$\frac{\partial}{\partial x_g^{(j)}} v_j(x^{(j)}) = \rho_g \left(\sum_{T \subseteq [M] \setminus \{j\}} c_T \mathcal{G}'(x_g^{(T \cup \{j\})}) \right) - \kappa_g. \tag{3.30}$$

We want to write seller j's marginal utility in a form that can be easily compared with seller k's marginal utility. We can accomplish this by changing the index set of the summation from $[M] \setminus \{j\}$ to $[M] \setminus \{j, k\}$ as follows,

$$\frac{\partial}{\partial x_g^{(j)}} v_j(x^{(j)}) = \rho_g \left(\sum_{T \subseteq [M] \setminus \{j,k\}} c_T \mathcal{G}'(x_g^{(T \cup \{j\})}) + c_{T \cup \{k\}} \mathcal{G}'(x_g^{((T \cup \{k\}) \cup \{j\})}) \right) - \kappa_g. \tag{3.31}$$

Similarly, write seller k's marginal utility with respect to $x_g^{(k)}$ as

$$\frac{\partial}{\partial x_g^{(k)}} v_k(x^{(k)}) = \rho_g \left(\sum_{T \subseteq [M] \setminus \{j,k\}} c_T \mathcal{G}'(x_g^{(T \cup \{k\})}) + c_{T \cup \{k\}} \mathcal{G}'(x_g^{((T \cup \{j\}) \cup \{k\})}) \right) - \kappa_g. \tag{3.32}$$

Note that $\mathcal{G}'$ is strictly decreasing. Without loss of generality, assume that $x_g^{(j)} > x_g^{(k)}$, then for every $T \subseteq [M] \setminus \{j, k\}$ it holds that

$$\mathcal{G}'(x_g^{(T \cup \{j\})}) < \mathcal{G}'(x_g^{(T \cup \{k\})}),$$

and

$$\mathcal{G}'(x_g^{((T \cup \{k\}) \cup \{j\})}) = \mathcal{G}'(x_g^{((T \cup \{j\}) \cup \{k\})}).$$

Consequently, the derivative of seller j's utility is strictly less than that of seller k's. We conclude

σ is not a Nash equilibrium. □

Fact (3.4.3) indicates that when the sellers face a common cost structure, the Shapley Value pushes the sellers to play the same strategies at equilibrium. This allows us to simplify the equilibrium equations to calculate for each group the exact amount of data that is produced when the sellers produce a positive amount of data.

Lemma 3.4.1. *(Baseline Data Production) If $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ is a Nash equilibrium at which samples are produced for some group g, i.e., $x_g^{([M])} > 0$, then every seller j produces $x_g^{(j)}$ samples given by,*

$$x_g^{(j)} = \frac{1}{M}\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{1/(\beta+1)}, \tag{3.33}$$

and the total number of samples produced over all the sellers is,

$$x_g^{([M])} = \left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{1/(\beta+1)}. \tag{3.34}$$

Proof. Seller j's utility is given by

$$v_j(x^{(j)}) = \left(\sum_{i=1}^{N}\sum_{g\in G}\mathcal{P}\mathcal{D}_{i,g,j}(x^{(j)})\right) - \kappa^{(j)T}x^{(j)} \tag{3.35}$$

$$= \sum_{g\in G}\rho_g \sum_{T\subseteq[M]\setminus\{j\}} c_T\left(\mathcal{G}(x_g^{(T\cup\{j\})}) - \mathcal{G}(x_g^{(T)})\right) - \sum_{g\in G}\kappa_g x_g^{(j)}. \tag{3.36}$$

By Fact (3.4.3) the sellers all make the same production decisions at equilibrium, i.e. for every pair of sellers $j, k \in [M]$ we have $x^{(j)} = x^{(k)}$. Consequently, every seller makes the same average marginal contribution over all the coalitions, so the sellers split the revenues collected from each buyer evenly. Therefore, seller j's utility is

$$v_j(x^{(j)}) = \frac{1}{M}\sum_{g\in G}\rho_g\mathcal{G}(x_g^{([M])}) - \kappa_g x_g^{(j)}. \tag{3.37}$$

Fact (3.4.3) also implies that $x_g^{([M])} = Mx_g^{(j)}$, thus

$$v_j(x^{(j)}) = \frac{1}{M}\sum_{g\in G}\rho_g \mathcal{G}(Mx_g^{(j)}) - \kappa_g x_g^{(j)}. \tag{3.38}$$

Seller j's marginal utility in $x_g^{(j)}$ is therefore,

$$\frac{\partial}{\partial x_g^{(j)}} v_j(x^{(j)}) = \frac{1}{M}\rho_g \left(\frac{\partial}{\partial x_g^{(j)}}\mathcal{G}(Mx_g^{(j)})\right) - \kappa_g. \tag{3.39}$$

Now,

$$\frac{\partial}{\partial x_g^{(j)}}\mathcal{G}(Mx_g^{(j)}) = \begin{cases} 0 & \text{if } Mx_g^{(j)} < (\alpha/Z)^{1/\beta} \\ \alpha\beta M^{-\beta}(x_g^{(j)})^{-\beta-1} & \text{if } (\alpha/Z)^{1/\beta} < Mx_g^{(j)} \end{cases} \tag{3.40}$$

If $Mx_g^{(j)} < (\alpha/Z)^{1/\beta}$, then seller j's marginal utility is negative. By assumption, σ is a Nash equilibrium at which the sellers produce data. At equilibrium, seller j's marginal utility in each $x_g^{(j)}$ must be 0. So we must have $Mx_g^{(j)} > (\alpha/Z)^{1/\beta}$, and seller j's marginal utility is

$$\frac{\partial}{\partial x_g^{(j)}} v_j(x^{(j)}) = \frac{1}{M}\rho_g \alpha\beta M^{-\beta}(x_g^{(j)})^{-\beta-1} - \kappa_g. \tag{3.41}$$

Setting this equal to 0 and solving for $x_g^{(j)}$ completes the proof. □

Lemma (3.4.1) indicates that when the sellers produce data, they produce more data for groups with greater potential economic value, ρ_g, and lower production costs, κ_g. Interestingly, Lemma (3.4.1) also indicates that the diminishing returns of increasing data dampens the effect that changes in potential economic value and production costs have on data production. Lemma (3.4.1) applies when the sellers produce data. But using it, we can characterize the conditions under which the sellers produce data. Before we do so, we prove a simple fact that we will apply in several analyses.

Fact 3.4.4. *If $\beta > 0$, then*

$$\beta\left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}}\right)^{\frac{\beta+1}{\beta}} > 1. \tag{3.42}$$

Proof. Towards contradition, suppose that

$$\beta \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}} \right)^{\frac{\beta+1}{\beta}} \leq 1. \tag{3.43}$$

And write,

$$\beta \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}} \right)^{\frac{\beta+1}{\beta}} \leq 1 \tag{3.44}$$

$$\implies \beta^{\frac{\beta}{\beta+1}} \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}} \right) \leq 1 \tag{3.45}$$

$$\implies 1 + \beta \leq 1 \tag{3.46}$$

$$\implies \beta \leq 0. \tag{3.47}$$

But $\beta > 0$ by definition of the model, a contradiction. □

Claim 3.4.1. *(Sellers' Baseline Participation Threshold) Let $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ be a Nash equilibrium, and g be any group in G. The sellers produce a positive number of samples for group g, i.e., $x_g^{([M])} > 0$, if and only if $\kappa_g \leq \tau_g$, where τ_g is a threshold value given by*

$$\tau_g \triangleq \rho_g c_{\mathcal{G}}, \tag{3.48}$$

where

$$c_{\mathcal{G}} \triangleq \frac{Z^{\frac{\beta+1}{\beta}}}{\alpha^{\frac{1}{\beta}} \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}} \right)^{\frac{\beta+1}{\beta}}}. \tag{3.49}$$

Proof. We first prove the forward direction, i.e., if $x_g^{([M])} > 0$, then $\kappa_g \leq \tau_g$. Fix a seller j. By assumption, σ is a Nash equilibrium, so every seller produces the same dataset at equilibrium by Fact (3.4.3). Hence seller j's utility is

$$v_j(x^{(j)}) = \frac{1}{M} \sum_{g \in G} \rho_g \mathcal{G}(M x_g^{(j)}) - \kappa_g x_g^{(j)}. \tag{3.50}$$

Also by assumption, $x_g^{([M])} > 0$. Since value is extracted independently by group in the baseline scenario, it must hold that

$$\frac{1}{M}\rho_g \mathcal{G}(M x_g^{(j)}) - \kappa_g x_g^{(j)} \geq 0, \tag{3.51}$$

because otherwise seller j could improve its utility by producing no data for group g and σ would not be a Nash equilibrium.

By Lemma (3.4.1),

$$x_g^{(j)} = \frac{1}{M}\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}}. \tag{3.52}$$

Therefore, we can write Inequality (3.51) as

$$\rho_g \mathcal{G}\left(\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}}\right) - \kappa_g\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}} \geq 0. \tag{3.53}$$

Recall that $\mathcal{G}$ is defined piecewise, we must ascertain which piece applies, i.e., whether or not,

$$\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}} \geq \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}. \tag{3.54}$$

Observe that if Inequality (3.54) does not hold, then seller j can improve its utility by producing no data for group g and σ would not be a Nash equilibrium. Therefore, Inequality (3.54) holds, and we have

$$\mathcal{G}\left(\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}}\right) = Z - \alpha\left(\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}}\right)^{-\beta}. \tag{3.55}$$

Substituting Equation (3.55) into Inequality (3.53) and some straightforward algebra yields:

$$\kappa_g \leq \frac{\rho_g Z^{\frac{\beta+1}{\beta}}}{\alpha^{\frac{1}{\beta}}\left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}}\right)^{\frac{\beta+1}{\beta}}} = \tau_g. \tag{3.56}$$

This proves the forward direction since j is arbitrary.

We now prove the reverse direction, i.e., if $\kappa_g \leq \tau_g$, then $x_g^{([M])} > 0$. We must show that the sellers will obtain non-negative utility by producing a positive number of samples, i.e., Inequality

(3.51) holds. To do so, we must evaluate

$$\mathcal{G}\left(\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}}\right) \tag{3.57}$$

which depends on whether Inequality (3.54) holds. We first show that it does.

By assumption, $\kappa_g \leq \tau_g$ and therefore

$$\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}} \geq \left(\frac{\rho_g}{\tau_g}\alpha\beta\right)^{\frac{1}{\beta+1}} \tag{3.58}$$

$$= \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}\left(\beta\left(\beta^{-\frac{\beta}{\beta+1}}+\beta^{\frac{1}{\beta+1}}\right)^{\frac{\beta+1}{\beta}}\right)^{\frac{1}{\beta+1}} \tag{3.59}$$

$$> \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}, \tag{3.60}$$

since

$$\beta\left(\beta^{-\frac{\beta}{\beta+1}}+\beta^{\frac{1}{\beta+1}}\right)^{\frac{\beta+1}{\beta}} > 1, \tag{3.61}$$

by Fact (3.4.4).

Thus, we can write Inequality (3.51) as

$$\frac{1}{M}\rho_g\left(Z-\alpha\left(\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}}\right)^{-\beta}\right)-\kappa_g\frac{1}{M}\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}} \geq 0. \tag{3.62}$$

Applying the assumption $\kappa_g \leq \tau_g$ and some straightforward algebra complete the proof. □

Claim (3.4.1) indicates that the seller's participation at Nash equilibrium depends critically on the relation between κ_g and τ_g. If $\kappa_g > \tau_g$, then the seller's utility will be negative for producing any amount of data. Therefore the seller will opt out of the data market for group g. If $\kappa_g = \tau_g$, then the seller is indifferent to participating because it obtains zero utility by opting out and zero utility at best by producing the utility-maximizing amount of data. Finally, if $\kappa_g < \tau_g$, then the seller will participate because producing the utility-maximizing amount of data will yield the seller positive utility.

Putting it all together, we can describe the baseline scenario equilibrium as follows.

Theorem 3.4.1. *If* $(p, \{\mu_i\}, \{x^{(j)}\})$ *is a Nash equilibrium, then for each group* g*, the marketplace sets price* p_g *to maximize* ρ_g *and the sellers produce a number of samples* $x_g^{(j)}$ *depending on the following inequality,*

$$\kappa_g \leq \tau_g. \tag{3.63}$$

If Inequality (3.63) holds, then for each group $g \in G$*, every seller* j *produces*

$$x_g^{(j)} = \frac{1}{M}\left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{1/(\beta+1)} \tag{3.64}$$

samples of group g*. If Inequality (3.63) does not hold, then* $x_g^{(j)} = 0$*.*

Theorem (3.4.1) enables us to describe the aggregate-dataset demographics that hold at equilibrium.

Corollary 3.4.1. *Let* $(p, \{\mu_i\}, \{x^{(j)}\})$ *be a Nash equilibrium, and* $H = \{g \in G : \kappa_g \leq \tau_g\}$*. Then for every group* g*, if* $g \in H$*, then,*

$$\frac{x_g^{([M])}}{\|x^{([M])}\|} = \frac{\left(\frac{\rho_g}{\kappa_g}\right)^{1/(\beta+1)}}{\sum_{h\in H}\left(\frac{\rho_h}{\kappa_h}\right)^{1/(\beta+1)}}, \tag{3.65}$$

and $\frac{x_g^{([M])}}{\|x^{([M])}\|} = 0$ *otherwise.*

Corollary (3.4.1) reveals how the aggregate-dataset demographics at equilibrium are driven by a complex interaction between the economics and the machine learning of the data market. Economic disparities, as captured in the quantities ρ_g/κ_g, drive the disparities in the aggregate-dataset demographics. But the effect is dampened towards uniformity by the diminishing gains in accuracy as the amount of training data grows as reflected by β. Yet the machine learning can also have the opposite effect on unfairness: it can amplify unfairness to the largest possible extent. The parameter α, which controls the scale of the excess errors, can render data production for some

groups unprofitable while allowing data production for other groups to be profitable. Groups for which data production is unprofitable will then have no representation in the aggregate dataset at equilibrium.

These findings illuminate the complex dynamics of the data market in the baseline scenario that can lead to unfair data production at equilibrium if left unchecked. They further demonstrate the need to understand the possibility of intervening for fairness which we study next.

3.4.2 Intervention Equilibrium

In the intervention scenario, the marketplace chooses a target vector γ that represents its assessment of which dataset demographics are fair and representative. The marketplace then intervenes to ensure that the datasets brought to market are γ-demographically balanced; it accepts a seller's dataset $x^{(j)}$ if and only if $x^{(j)}$ is γ-demographically balanced. We now investigate the equilibrium strategies of the agents in the intervention scenario.

Our previous analyses have already established the equilibrium strategies of the marketplace and sellers. We now analyze the sellers' equilibrium strategies. As in the baseline scenario, the sellers will produce the same datasets at equilibrium because they face a common cost structure.

Fact 3.4.5. *(Symmetric Sellers Intervention Equilibrium Strategies) If* $(p, \{\mu_i\}, \{x^{(j)}\})$ *is a Nash equilibrium, then for every* $j, k \in [M]$ *we have that* $x^{(j)} = x^{(k)}$.

Proof. In the intervention scenario, the marketplace's intervention couples data production across the groups. Therefore, the production decision that each seller j faces is how many samples in total to produce, i.e., $n^{(j)} \triangleq \|x^{(j)}\|$, because the number of samples of each group, $x_g^{(j)}$, is then determined by the marketplace's choice of target vector γ, i.e. $x_g^{(j)} = \gamma_g n^{(j)}$. Define $n^{(T)} \triangleq \sum_{k\in T} n^{(k)}$ for any $T \subseteq [M]$ and write seller j's utility by

$$v_j(x^{(j)}) = \sum_{g\in G} p_g \sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \sum_{T\subseteq[M]\setminus\{j\}} c_T \left(\mathcal{G}(x_g^{(T\cup\{j\})}) - \mathcal{G}(x_g^{(T)})\right) - \sum_{g\in G} \kappa_g x_g^{(j)} \tag{3.66}$$

$$= \sum_{g\in G} \rho_g \sum_{T\subseteq[M]\setminus\{j\}} c_T \left(\mathcal{G}(\gamma_g n^{(T\cup\{j\})}) - \mathcal{G}(\gamma_g n^{(T)})\right) - \kappa^T \gamma n^{(j)} \triangleq v_j(n^{(j)}). \tag{3.67}$$

At equilibrium, every seller j's marginal utility must be 0. Therefore, for any two sellers j and k, seller j's marginal utility must equal seller k's marginal utility; formally, we must have,

$$\frac{\partial}{\partial n^{(j)}} v_j(n^{(j)}) = 0 = \frac{\partial}{\partial n^{(k)}} v_k(n^{(k)}). \tag{3.68}$$

We now show that if two sellers, j and k produce a different total number of samples, i.e., $n^{(j)} \neq n^{(k)}$, then $(p, \{\mu_i\}, \{x^{(j)}\})$ is not a Nash equilibrium. Write seller j's marginal utility with respect to $n^{(j)}$ as

$$\frac{\partial}{\partial n^{(j)}} v_j(n^{(j)}) = \sum_{g \in G} \rho_g \left(\sum_{T \subseteq [M] \setminus \{j,k\}} c_T \mathcal{G}'(\gamma_g n^{(T \cup \{j\})}) + c_{T \cup \{k\}} \mathcal{G}'(\gamma_g n^{((T \cup \{k\}) \cup \{j\})}) \right) - \kappa^T \gamma. \tag{3.69}$$

Write seller k's marginal utility with respect to $n^{(k)}$ as

$$\frac{\partial}{\partial n^{(k)}} v_k(n^{(k)}) = \rho_g \left(\sum_{T \subseteq [M] \setminus \{j,k\}} c_T \mathcal{G}'(\gamma_g n^{(T \cup \{k\})}) + c_{T \cup \{k\}} \mathcal{G}'(\gamma_g n^{((T \cup \{j\}) \cup \{k\})}) \right) - \kappa^T \gamma. \tag{3.70}$$

Note that $\mathcal{G}'$ is strictly decreasing. Without loss of generality and towards contradiction, assume that $n^{(j)} > n^{(k)}$, then for every $T \subseteq [M] \setminus \{j, k\}$ it holds that

$$\mathcal{G}'(\gamma_g n^{(T \cup \{j\})}) < \mathcal{G}'(\gamma_g n^{(T \cup \{k\})}), \tag{3.71}$$

and

$$\mathcal{G}'(\gamma_g n^{((T \cup \{k\}) \cup \{j\})}) = \mathcal{G}'(\gamma_g n^{((T \cup \{j\}) \cup \{k\})}). \tag{3.72}$$

Consequently, the derivative of seller j's utility is strictly less than that of seller k's, a contradiction. We conclude that $n^{(j)} = n^{(k)}$, i.e., $x^{(j)} = x^{(k)}$. □

Fact (3.4.5) again indicates that when the sellers face a common cost structure, the Shapley Value pushes the sellers to play the same strategies at equilibrium. This again allows us to simplify the equilibrium equations to calculate the amount of data the sellers produce when they produce

data. But unlike in the baseline scenario, the sellers do not decide how many training samples to produce for each group independently, the proportions are determined by γ. Thus, the sellers production decision can be described in terms of the total number of samples they decide to produce.

Lemma 3.4.2. *(Intervention Data Production) Fix a target vector γ. If $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ is a Nash equilibrium such that samples are produced, i.e. $n^{([M])} > 0$, then there exists a group $h \in G$ such that every seller j produces*

$$n^{(j)} = \frac{1}{M}\left(\frac{\alpha\beta}{\kappa^T\gamma}\sum_{g\in H}\rho_g\gamma_g^{-\beta}\right)^{1/(\beta+1)} \tag{3.73}$$

samples, where

$$H \triangleq \{g \in G : \gamma_g \geq \gamma_h\}, \tag{3.74}$$

and γ_h is a minimum value over γ_g satisfying

$$M\gamma_g n^{(j)} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}. \tag{3.75}$$

Proof. Write seller j's utility as follows

$$v_j(x^{(j)}) = \left(\sum_{i=1}^{N} \sum_{g \in G} \mathcal{PD}_{i,g,j}(x^{(j)}) \right) - \kappa^{(j)T} x^{(j)} \tag{3.76}$$

$$= \left(\sum_{i=1}^{N} \sum_{g \in G} p_g \sum_{T \subseteq [M] \setminus \{j\}} c_T \cdot \left(\mathcal{G}_{i,g}(\mathcal{AF}_g(\mu_{i,g}, x^{(T \cup \{j\})})) - \mathcal{G}_{i,g}(\mathcal{AF}_g(\mu_{i,g}, x^{(T)})) \right) \right) - \kappa^{(j)T} x^{(j)} \tag{3.77}$$

$$= \left(\sum_{i=1}^{N} \sum_{g \in G} p_g \frac{1}{M} \mathcal{G}_{i,g}(\mathcal{AF}_g(\mu_{i,g}, x^{([M])})) \right) - \kappa^{(j)T} x^{(j)} \tag{3.78}$$

$$= \left(\sum_{i=1}^{N} \sum_{g \in G} p_g \mathbf{1}[\mu_{i,g} \geq p_g] \frac{1}{M} \mathcal{G}_{i,g}(x^{([M])}) \right) - \kappa^{(j)T} x^{(j)} \tag{3.79}$$

$$= \left(\sum_{i=1}^{N} \sum_{g \in G} p_g \mathbf{1}[\mu_{i,g} \geq p_g] \frac{1}{M} \mathcal{G}(x^{([M])}) \right) - \kappa^{T} x^{(j)} \tag{3.80}$$

$$= \left(\sum_{i=1}^{N} \sum_{g \in G} p_g \mathbf{1}[\mu_{i,g} \geq p_g] \frac{1}{M} \mathcal{G}(x_g^{([M])}) \right) - \kappa^{T} x^{(j)} \tag{3.81}$$

$$= \frac{1}{M} \left(\sum_{g \in G} p_g \left(\sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \right) \mathcal{G}(x_g^{([M])}) \right) - \kappa^{T} x^{(j)} \tag{3.82}$$

$$= \frac{1}{M} \left(\sum_{g \in G} \rho_g \mathcal{G}(x_g^{([M])}) \right) - \kappa^{T} x^{(j)}. \tag{3.83}$$

In order: Equation (3.76) is by definition of the seller's utility; Equation (3.77) is by definition of the payment division function; Equation (3.78) is by Fact (3.4.5) since the sellers all play the same strategy; Equation (3.79) is by definition of the allocation function; Equation (3.80) is by quasi-symmetry; Equation (3.81) is by the assumption of zero group inter-transfer; Equation (3.82) is by rearranging terms; and Equation (3.83) is by definition of market value-of-accuracy.

In the intervention scenario, the marketplace's target vector, γ, determines the sellers' marginal production costs as $\kappa^T \gamma$, and each participating seller j has only to decide how many samples to produce, $n^{(j)}$, incurring total production costs of $\kappa^T \gamma n^{(j)}$. And again, because all the seller's

produce the same number of samples at equilibrium, seller j's utility becomes

$$v_j(n^{(j)}) = \frac{1}{M}\left(\sum_{g\in G}\rho_g \mathcal{G}(M\gamma_g n^{(j)})\right) - \kappa^T \gamma n^{(j)}. \tag{3.84}$$

By assumption, the sellers produce some samples, i.e., $n^{([M])} > 0$, and so $n^{(j)} > 0$. It follows that there must exist at least one group $g \in G$ satisfying

$$M\gamma_h n^{(j)} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}, \tag{3.85}$$

because otherwise seller j's utility would be negative and σ would not be a Nash equilibrium.

Now, let h be the group with the smallest γ_h satisfying Inequality (3.85). Then, for every $g \in G$ satisfying $\gamma_g \geq \gamma_h$ we have

$$M\gamma_g n^{(j)} \geq M\gamma_h n^{(j)} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}. \tag{3.86}$$

Therefore we can define

$$H \triangleq \{g \in G : \gamma_g \geq \gamma_h\}, \tag{3.87}$$

and seller j's utility becomes

$$v_j(n^{(j)}) = \frac{1}{M}\left(\sum_{g\in H}\rho_g \left(Z - \alpha(M\gamma_g n^{(j)})^{-\beta}\right)\right) - \kappa^T \gamma n^{(j)}. \tag{3.88}$$

Therefore, seller j's marginal utility in $n^{(j)}$ is

$$v'(n^{(j)}) = \frac{1}{M}\sum_{g\in H}\rho_g \alpha\beta \left(M\gamma_g n^{(j)}\right)^{-\beta-1} M\gamma_g - \kappa^T\gamma. \tag{3.89}$$

At equilibrium, the seller's marginal utility is 0; solving for $n^{(j)}$ completes the proof. □

Lemma (3.4.2) shows how coupling data production across the groups via γ affects data production. The sellers produce more data as the group's market value-of-accuracy increases, but this is mediated by their required representation, $\rho_g \gamma_g^{-\beta}$. And the sellers produce more data as the

marginal production cost $\kappa^T \gamma$ decreases. We can use Lemma (3.4.2) to characterize the conditions under which the sellers produce no data and dispel this bizarre seeming-possibility.

Claim 3.4.2. *(Sellers' Intervention Participation Thresholds) Fix a target vector γ. Let $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ be a Nash equilibrium. If the sellers produce a positive number of samples, i.e., $n^{([M])} > 0$, then there exists a group $h \in G$ such that the sellers' marginal production cost, $\kappa^T \gamma$, is at most a threshold value, $\tau_H(\rho, \gamma)$, given by,*

$$\tau_H(\rho, \gamma) = \frac{\left(\sum_{g \in H} \rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(\sum_{g \in H} \rho_g \gamma_g^{-\beta}\right)^{\frac{1}{\beta}}} \cdot c_{\mathcal{G}} \tag{3.90}$$

where

$$H \triangleq \{g \in G : \gamma_g \geq \gamma_h\}, \tag{3.91}$$

and γ_h is a minimum value over γ_g satisfying

$$\gamma_g n^{([M])} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}. \tag{3.92}$$

Proof. Fix a seller j, and write its utility as follows

$$v_j(x^{(j)}) = \left(\sum_{i=1}^{N}\sum_{g\in G}\mathcal{PD}_{i,g,j}(x^{(j)})\right) - \kappa^{(j)T}x^{(j)} \tag{3.93}$$

$$= \left(\sum_{i=1}^{N}\sum_{g\in G} p_g \sum_{T\subseteq[M]\setminus\{j\}} c_T \cdot \left(\mathcal{G}_{i,g}(\mathcal{AF}_g(\mu_{i,g}, x^{(T\cup\{j\})})) - \mathcal{G}_{i,g}(\mathcal{AF}_g(\mu_{i,g}, x^{(T)}))\right)\right) - \kappa^{(j)T}x^{(j)} \tag{3.94}$$

$$= \left(\sum_{i=1}^{N}\sum_{g\in G} p_g \frac{1}{M}\mathcal{G}_{i,g}(\mathcal{AF}_g(\mu_{i,g}, x^{([M])}))\right) - \kappa^{(j)T}x^{(j)} \tag{3.95}$$

$$= \left(\sum_{i=1}^{N}\sum_{g\in G} p_g \mathbf{1}[\mu_{i,g} \geq p_g]\frac{1}{M}\mathcal{G}_{i,g}(x^{([M])})\right) - \kappa^{(j)T}x^{(j)} \tag{3.96}$$

$$= \left(\sum_{i=1}^{N}\sum_{g\in G} p_g \mathbf{1}[\mu_{i,g} \geq p_g]\frac{1}{M}\mathcal{G}(x^{([M])})\right) - \kappa^{T}x^{(j)} \tag{3.97}$$

$$= \left(\sum_{i=1}^{N}\sum_{g\in G} p_g \mathbf{1}[\mu_{i,g} \geq p_g]\frac{1}{M}\mathcal{G}(x_g^{([M])})\right) - \kappa^{T}x^{(j)} \tag{3.98}$$

$$= \frac{1}{M}\left(\sum_{g\in G} p_g \left(\sum_{i=1}^{N}\mathbf{1}[\mu_{i,g} \geq p_g]\right)\mathcal{G}(x_g^{([M])})\right) - \kappa^{T}x^{(j)} \tag{3.99}$$

$$= \frac{1}{M}\left(\sum_{g\in G}\rho_g\mathcal{G}(x_g^{([M])})\right) - \kappa^{T}x^{(j)}. \tag{3.100}$$

In order: Equation (3.93) is by definition of the seller's utility; Equation (3.94) is by definition of the payment division function; Equation (3.95) is by Fact (3.4.5) since the sellers all play the same strategy; Equation (3.96) is by definition of the allocation function; Equation (3.97) is by quasi-symmetry; Equation (3.98) is by the assumption of zero group inter-transfer; Equation (3.99) is by rearranging terms; and Equation (3.100) is by definition of economic potential.

In the intervention scenario, the marketplace's target vector, γ, determines the sellers' marginal production costs as $\kappa^T\gamma$, and each participating seller j has only to decide how many samples to produce, $n^{(j)}$, incurring total production costs of $\kappa^T\gamma n^{(j)}$. And again, because all the seller's

produce the same number of samples at equilibrium, seller j's utility becomes

$$v_j(n^{(j)}) = \frac{1}{M}\left(\sum_{g\in G}\rho_g \mathcal{G}(M\gamma_g n^{(j)})\right) - \kappa^T \gamma n^{(j)}. \tag{3.101}$$

By Lemma (3.4.2), there exists a group $h \in G$ such that every seller j produces

$$n^{(j)} = \frac{1}{M}\left(\frac{\alpha\beta}{\kappa^T\gamma}\sum_{g\in H}\rho_g\gamma_g^{-\beta}\right)^{1/(\beta+1)} \tag{3.102}$$

samples, where

$$H \triangleq \{g \in G : \gamma_g \geq \gamma_h\}, \tag{3.103}$$

and γ_h is a minimum value over γ_g satisfying

$$M\gamma_g n^{(j)} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}. \tag{3.104}$$

Therefore, seller j's utility is

$$v_j(n^{(j)}) = \frac{1}{M}\left(\sum_{g\in H}\rho_g \mathcal{G}(M\gamma_g n^{(j)})\right) - \kappa^T\gamma n^{(j)} \tag{3.105}$$

$$= \frac{1}{M}\left(\sum_{g\in H}\rho_g\left(Z - \alpha\left(M\gamma_g n^{(j)}\right)^{-\beta}\right)\right) - \kappa^T\gamma n^{(j)} \tag{3.106}$$

$$= \frac{1}{M}\left(\sum_{g\in H}\rho_g\left(Z - \alpha\left(M\gamma_g\frac{1}{M}\left(\frac{\alpha\beta}{\kappa^T\gamma}\sum_{f\in H}\rho_f\gamma_f^{-\beta}\right)^{\frac{1}{\beta+1}}\right)^{-\beta}\right)\right) - \kappa^T\gamma\frac{1}{M}\left(\frac{\alpha\beta}{\kappa^T\gamma}\sum_{f\in H}\rho_f\gamma_f^{-\beta}\right)^{\frac{1}{\beta+1}} \tag{3.107}$$

By assumption, σ is a Nash equilibrium, hence seller j's utility is non-negative; solving for $\kappa^T\gamma$ with straightforward algebra completes the proof. □

Claim (3.4.2) tells us that if the sellers produce data at Nash equilibrium in the intervention scenario, then they must be profitably monetizing some groups. This parallels Claim (3.4.1), which

tells us that if the sellers produce data for a group at Nash equilibrium in the baseline scenario, then they must be profitably monetizing that group. But there is a notable divergence between the two scenarios. In the baseline scenario, knowing that a group can be profitably monetized also determines the amount of data produced for that group at Nash equilibrium. In the intervention scenario, however, knowing that some subset of groups can be profitably monetized does not by itself determine the amount of data being produced because there may exist another subset of groups that can be monetized more profitably. Altogether, we have the following description of the intervention scenario equilibrium.

Theorem 3.4.2. *Let $(p, \{\mu_i\}, \{x^{(j)}\})$ be a Nash equilibrium, then for each group g, the marketplace sets reserve price p_g to maximize ρ_g and each seller j produces data if and only if the sellers' marginal production cost, $\kappa^T \gamma$, is at most a threshold value, $\tau_H(\rho, \gamma)$, given by,*

$$\tau_H(\rho, \gamma) = \frac{\left(\sum_{g \in H} \rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(\sum_{g \in H} \rho_g \gamma_g^{-\beta}\right)^{\frac{1}{\beta}}} \cdot c_{\mathcal{G}} \tag{3.108}$$

where

$$H \triangleq \{g \in G : \gamma_g \geq \gamma_h\}, \tag{3.109}$$

for some group h such that γ_h is a minimum value over γ_g satisfying

$$\gamma_g n^{([M])} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}. \tag{3.110}$$

Theorems (3.4.1) and (3.4.2) describe the data market equilibria in the baseline and intervention scenarios, respectively. In the next section, we derive insights into problem of intervening for fairness in the model.

3.5 Intervening for fairness can backfire

In order to build fair machine-learning systems, practitioners must typically be willing to incur some costs. While the fairness benefits may justify those costs, it is still important that any fairness intervention impose an acceptable cost and achieve the intended fairness benefits. What is an acceptable cost is debatable, but an intervention that imposes any cost and does not obtain any fairness benefits—or even harms the groups it is intended to benefit—is categorically unacceptable. When this happens despite the best intentions of the intervener, we say that the intervention backfires. In this section we study the conditions under which an intervention backfires in the model.

Definition 3.5.1. *(Intervention backfire) Let γ be a target vector, and fix a data market. We say that* the intervention γ backfires in the data market *if at least one group-specific sub-market forms in the baseline scenario but the market does not form in the intervention scenario.*

Our first result shows that there is a risk of interventions backfiring.

Theorem 3.5.1. *For every target vector γ there exists a data market in which γ backfires.*

Proof. Fix a target vector γ. We will construct a data market in which γ backfires. Let the N buyers be arbitrary. Let there be M sellers whose common cost structure κ is arbitrary except for two groups h and h' that we will specify. We will set κ_h so that the h-specific sub-market forms in the baseline scenario, and we will set $\kappa_{h'}$ so that the market does not form in the intervention scenario. In other words, we will show that: 1) in the baseline scenario there exists a Nash equilibrium $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ such that $x_h^{([M])} > 0$; and 2) in the intervention scenario there does not exist a Nash equilibrium $\sigma' = (p, \{\mu_i\}, \{y^{(j)}\})$ such that $\|y^{([M])}\| > 0$.

We show 1) first. By Fact (3.4.2), the marketplace sets the reserve prices p_g to maximize ρ_g for every group g in both scenarios. Now set κ_h to any value satisfying $\kappa_h \leq \tau_h$. We will set $\kappa_{h'}$ more precisely when we turn to the intervention scenario, but we require it to satisfy $\kappa_{h'} > \tau_{h'}$. For every seller j, and group g set $x_g^{(j)}$ in accordance with Theorem (3.4.1). It follows that $\sigma =$

$(p, \{\mu_i\}, \{x^{(j)}\})$ is a Nash equilibrium in the baseline scenario such that $x_h^{([M])} > 0$. We conclude that the data market forms in the baseline scenario.

We now show 2). We must further specify $\kappa_{h'}$. We wish to ensure that $\kappa^T \gamma > \tau_H(\rho, \gamma)$ for every $H \subseteq G$. Observe that once the buyers are fixed, then the maximum value of $\tau_H(\rho, \gamma)$ over all possible H is determined and finite. Define

$$\lceil \tau \rceil \triangleq \max_{H \subseteq G} \tau_H(\rho, \gamma). \tag{3.111}$$

Now we just need to ensure that

$$\kappa^T \gamma > \lceil \tau \rceil, \tag{3.112}$$

which readily follows by setting $\kappa_{h'}$ to any value satisfying,

$$\kappa_{h'} > \frac{\lceil \tau \rceil}{\gamma_{h'}}. \tag{3.113}$$

Thus, the sellers' intervention marginal cost of production is greater than intervention participation threshold over all possible H and 2) follows. We conclude that the data market does not form in the intervention scenario. □

Theorem (3.5.1) tells us that all interventions are risky when the intervention requires the sellers to begin producing data for some group. The sellers produce no data in the baseline scenario because the group's production costs are prohibitive, and if they are sufficiently high, they can also be prohibitive in the intervention scenario. If the marketplace does not choose the target vector γ carefully, then the sellers may opt out of the market. This is striking because this magnifies the cost of fairness *to the maximum extent possible and to every single agent*, i.e., the cost of fairness for every agent is its full baseline utility. And this can harm the very groups that are the intended beneficiaries of the intervention. Some groups may be better off with some samples in the baseline scenario—even if they are under-represented—versus no samples in the intervention scenario.

Yet, Theorem (3.5.1) is also tempered by its assumption that no data is produced for some of the

groups in the baseline scenario. A regulator might consider addressing data production for those groups via alternative intervention strategies, e.g. subsidies or NGO grants, and focus its market intervention on groups for which data are produced in the baseline scenario. In the model, this would correspond to an intervention in a data market that fully forms in the baseline scenario and where the intervention stipulates some representation for each group. Such a divide-and-conquer approach might make sense if the backfire risk then behaves more manageably. Our next result shows that this can be an effective strategy for mitigating the backfire risk, at least for one specific target vector.

Definition 3.5.2. *(Uniform intervention) The* uniform intervention*, denoted u, is the target vector defined by*

$$u_g \triangleq \frac{1}{|G|}, \tag{3.114}$$

for every group g.

Theorem 3.5.2. *Fix a data market that is fully forming in the baseline scenario. If the marketplace chooses the uniform intervention, i.e., $\gamma = u$, then the data market forms in the intervention scenario.*

Proof. Fix N buyers and M sellers such that every group-specific sub-market forms at the Nash equilibrium, $(p, \{\mu_i\}, \{x^{(j)}\})$, in the baseline scenario. Because the market fully forms in the baseline scenario, the groups' production costs must not be too large. In particular, for each group g, its production costs, κ_g, must be at most the seller's baseline participation threshold, τ_g,

$$\kappa_g \leq \tau_g. \tag{3.115}$$

We want to show that the market forms in the intervention scenario, i.e., the sellers will produce data when the marketplace applies its fairness intervention with target vector u. The market forms when the seller's intervention production cost, $\kappa^T u$, is no more than the seller's intervention participation threshold, τ. The seller's intervention production cost is a function of κ and u, hence,

it is fixed once the sellers are fixed. In contrast, the seller's intervention participation threshold, τ, depends on the equilibrium strategies of the marketplace and the buyers in the intervention scenario.

Let $(p, \{\mu_i\}, \{y^{(j)}\})$ be a Nash equilibrium in the intervention scenario. We must show that there exists some subset of groups H such that $\kappa^T u \leq \tau_H(\rho, u)$ and for every $g \in H$, $n^{([M])}/M > (\alpha/Z)^{1/\beta}$. In fact, we will show that this holds for $H = G$. First we analyze $\tau_G(\rho, u)$,

$$\tau_G(\rho,u) = \frac{\left(\sum_{g\in G}\rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(\sum_{g\in G}\rho_g\gamma_g^{-\beta}\right)^{\frac{1}{\beta}}}c_{\mathcal{G}} = \frac{\left(\sum_{g\in G}\rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(\sum_{g\in G}\rho_g\left(\frac{1}{|G|}\right)^{-\beta}\right)^{\frac{1}{\beta}}}c_{\mathcal{G}} = \frac{\left(\sum_{g\in G}\rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(\sum_{g\in G}\rho_g|G|^{\beta}\right)^{\frac{1}{\beta}}}c_{\mathcal{G}} \tag{3.116}$$

$$= \frac{\left(\sum_{g\in G}\rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(|G|^{\beta}\sum_{g\in G}\rho_g\right)^{\frac{1}{\beta}}}c_{\mathcal{G}} = \frac{\left(\sum_{g\in G}\rho_g\right)^{\frac{\beta+1}{\beta}}}{|G|\left(\sum_{g\in G}\rho_g\right)^{\frac{1}{\beta}}}c_{\mathcal{G}} = \frac{1}{|G|}\sum_{g\in G}\rho_g c_{\mathcal{G}} \tag{3.117}$$

$$= \frac{1}{|G|}\sum_{g\in G}\tau_g. \tag{3.118}$$

By assumption, $\kappa_g \leq \tau_g$ for every group g. It follows that

$$\kappa^T u = \frac{1}{|G|}\sum_{g\in G}\kappa_g \leq \frac{1}{|G|}\sum_{g\in G}\tau_g = \tau_G(\rho,u). \tag{3.119}$$

It remains to show that

$$y_g^{([M])} = \frac{n^{([M])}}{M} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}. \tag{3.120}$$

When $H = G$ we have,

$$n^{([M])} = \left(\frac{\alpha\beta}{\kappa^T u}\sum_{g\in G}\rho_g u_g^{-\beta}\right)^{\frac{1}{\beta+1}} = \left(\frac{\alpha\beta}{\sum_{g\in G}\kappa_g\frac{1}{M}}\sum_{g\in G}\rho_g\left(\frac{1}{M}\right)^{-\beta}\right)^{\frac{1}{\beta+1}} = M\left(\alpha\beta\frac{\sum_{g\in G}\rho_g}{\sum_{g\in G}\kappa_g}\right)^{\frac{1}{\beta+1}} \tag{3.121}$$

$$\geq M\left(\frac{\alpha\beta}{c_{\mathcal{G}}}\right) > M\left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}. \tag{3.122}$$

This completes the proof. □

Theorem 3.5.3. *Let γ be the marketplace's target vector. If γ is not the uniform intervention, i.e., $\gamma \neq u$, then there exists a data-market instance $\mathcal{M}$ that is fully forming in the baseline scenario but does not form in the intervention scenario.*

Let us first discuss the proof. The key high-level idea is the following. We will construct a specific data market in which every $\gamma \neq u$ backfires.[6] This will be due to two key features of the data market: 1) it is just on the verge of formation; and 2) the dataset demographics at baseline equilibrium will be uniform. Thus, any γ that forces the sellers to deviate from their baseline equilibrium strategies also forces the sellers out of the data market. Although the result is intuitive, the proof is quite involved due to one principal technical challenge.

For any $\gamma \neq u$, it is straightforward to check that if the marketplace sets the same reserve prices in the intervention scenario as in the baseline-scenario equilibrium, then the market will not form in the intervention scenario. This observation proves that γ backfires when the marketplace does not change its reserve prices. The technical difficulty stems from the fact that the marketplace can, in principle, set different reserve prices in the intervention scenario. And Claim (3.4.2) indicates that sometimes the marketplace can increase the sellers' intervention participation threshold, τ, *by decreasing* some of its reserve prices. To prove the theorem requires us to prove that γ backfires for any set of reserve prices the marketplace may choose, not just the same ones as at baseline equilibrium. And this must be over all possible choices of γ.

The proof establishes the theorem by showing that over all possible γ and all possible reserve prices, the sellers' intervention participation threshold is less than the seller's marginal cost of production when $\gamma \neq u$. This is done in two major steps. In the first step, the quantification over all possible γ and reserve prices is reduced to a quantification over all possible reserve prices by computing the γ that maximizes the sellers' intervention participation threshold given a fixed set of reserve prices. And the second step proves that the maximum sellers' intervention participation threshold over all reserve prices is obtained when $\gamma = u$. This proves the theorem, because that is precisely the seller's marginal cost of production.

[6] We analyze a specific data market for ease and clarity of presentation, our analysis readily generalizes to a more restricted class.

Proof. We first specify the data market. There is only 1 buyer whose value-of-accuracy is the same across all the groups, i.e., for all $g \in G$, we have $\mu_{1,g} = c_\mu$ for some positive constant $c_\mu > 0$. There are M sellers that face a cost-structure κ, with the following two properties: 1) the marginal production cost is the same across all the groups, i.e., for all $g \in G$, we have $\kappa_g = c_\kappa$ for some positive constant $c_\kappa > 0$; and 2) c_κ is related to the buyer's value-of-accuracy and prediction gain function by $c_\kappa = c_\mu c_{\mathcal{G}}$.

We now analyze the baseline equilibrium. Since there is only one buyer, the marketplace's value-of-accuracy for group g when it plays reserve price p_g is p_g if $p_g \leq c_\mu$ and 0 otherwise. Therefore, the marketplace will set each group's reserve price to the buyer's value-of-accuracy and all the reserve prices will be the same, i.e., for all $g \in G$, $p_g = c_\mu$. Consequently, the seller's baseline participation threshold for each group g is $\tau_g = c_\mu c_{\mathcal{G}}$. By construction, for every group g, $\kappa_g = c_\kappa = c_\mu c_{\mathcal{G}}$, therefore $\kappa_g = \tau_g$ and every seller will produce samples of group g at equilibrium, i.e., $x_g^{(j)} > 0$. Moreover, every seller will produce the same number of samples for every group, i.e., for all $g, h \in G$, $x_g^{(j)} = x_h^{(j)}$. Therefore every group will have the same number of samples in the aggregate dataset, i.e., for all $g, h \in G$, $x_g^{([M])} = x_h^{([M])}$, and the demographics of the aggregate dataset will coincide with the uniform intervention, i.e., $\gamma(x^{([M])}) = u$.

We turn to analyzing the intervention scenario. We show that over all the marketplace's possible choices of target vector, γ, and reserve prices q, the sellers' intervention participation threshold reaches its maximum when $\gamma = u$ and $q_g = c_\mu$. We do this in two major steps. In the first step, we solve for the target vector γ that maximizes the seller's intervention participation threshold given fixed reserve prices q. This allows us to maximize the sellers' intervention participation threshold solely in terms of the reserve prices. In the second step, we show that the reserve prices that maximize the seller's intervention participation threshold are $q_g = c_\mu$, and this implies that $\gamma = u$.

We now take the first step. Fix the marketplace's reserve prices q in the intervention scenario.

The sellers' intervention participation threshold is

$$\tau = \frac{\left(\sum_{g\in H}\rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(\sum_{g\in H}\rho_g\gamma_g^{-\beta}\right)^{\frac{1}{\beta}}}\cdot c_{\mathcal{G}}. \tag{3.123}$$

Note that τ depends on q through ρ_g, and that the ρ_g are fixed because q is fixed. Therefore, choosing γ to maximize τ is equivalent to choosing γ to minimize

$$\sum_{g\in H}\rho_g\gamma_g^{-\beta}. \tag{3.124}$$

Define

$$f(\gamma) \triangleq \begin{cases} \sum_{g\in H}\rho_g\gamma_g^{-\beta} & \text{if } \forall h\in H, \gamma_h > 0 \\ \infty & \text{otherwise} \end{cases} \tag{3.125}$$

Thus, we wish to solve the following program:

$$\min_{\gamma} f(\gamma), \tag{3.126}$$

subject to

$$\sum_{g\in G}\gamma_g = 1, \tag{3.127}$$

and for all $g \in G$, $0 \le \gamma_g \le 1$.

Define the following functions for the constraints:

$$h(\gamma) \triangleq \sum_{g\in G}\gamma_g - 1; \tag{3.128}$$

and for every $g \in G$,

$$b_{(\ell,g)}(\gamma) \triangleq -\gamma_g, \tag{3.129}$$

and

$$b_{(u,g)}(\gamma) \triangleq \gamma_g - 1. \tag{3.130}$$

Compute the partial derivatives of the objective and constraint functions. For the objective,

$$\frac{\partial}{\partial \gamma_g} f(\gamma) = \begin{cases} -\beta \rho_g \gamma_g^{-\beta-1} & \text{if } g \in H \\ 0 & \text{otherwise} \end{cases} \tag{3.131}$$

For the equality constraint,

$$\frac{\partial}{\partial \gamma_g} h(\gamma) = 1. \tag{3.132}$$

For the lower bound constraints,

$$\frac{\partial}{\partial \gamma_h} b_{(\ell,g)}(\gamma) = \begin{cases} -1 & \text{if } g = h \\ 0 & \text{otherwise} \end{cases} \tag{3.133}$$

And for the upper bound constraints,

$$\frac{\partial}{\partial \gamma_h} b_{(u,g)}(\gamma) = \begin{cases} 1 & \text{if } g = h \\ 0 & \text{otherwise} \end{cases} \tag{3.134}$$

By the Karush-Kuhn-Tucker (KKT) conditions, we are searching for solutions γ that satisfy the multiplier rule,

$$\nabla f(\gamma) + \nabla b(\gamma)\lambda + \nabla h(\gamma)\mu = 0, \tag{3.135}$$

and complementarity conditions, i.e., $\lambda \geq 0$ and

$$b(\gamma)^T \lambda = 0. \tag{3.136}$$

Plug the partial derivatives into the KKT multiplier rule. For each $g \in H$ this gives

$$-\beta\rho_g\gamma_g^{-\beta-1} - \lambda_{(\ell,g)} + \lambda_{(u,g)} + \mu = 0. \tag{3.137}$$

And for each $g' \in G \setminus H$ this gives

$$-\lambda_{(\ell,g')} + \lambda_{(u,g')} + \mu = 0. \tag{3.138}$$

Now analyze the multipliers. By the definition of f, observe that an optimal solution γ must satisfy $\gamma_g > 0$ for all $g \in H$. This has a number of consequences. First, if $g \in H$, then the lower bound constraint is loose, i.e., $b_{(\ell,g)}(\gamma) < 0$, and the complementarity conditions imply that $\lambda_{(\ell,g)} = 0$. Second, if $g' \in G \setminus H$, then the upper bound constraint for g' is loose, i.e., $\gamma_{g'} < 1$, and the complementarity conditions imply that $\lambda_{(u,g')} = 0$. Finally, if $g' \in G \setminus H$, then $\gamma_{g'} = 0$, which can be seen as follows. Towards contradiction, suppose $\gamma_{g'} > 0$. Then the KKT multiplier rule for g' is

$$\mu = 0, \tag{3.139}$$

and the KKT multiplier rule for any $g \in H$ is

$$-\beta\rho_g\gamma_g^{-\beta-1} + \mu = -\beta\rho_g\gamma_g^{-\beta-1} + 0 = -\beta\rho_g\gamma_g^{-\beta-1} = 0. \tag{3.140}$$

But this is a contradiction because $\gamma_g > 0$ implies that $-\beta\rho_g\gamma_g^{-\beta-1} < 0$.

It remains to solve for γ_g, for each $g \in H$. If $|H| = 1$, it follows that $\gamma_g = 1$. Otherwise, $|H| > 1$, and the KKT multiplier rule is

$$-\beta\rho_g\gamma_g^{-\beta+1} + \mu = 0. \tag{3.141}$$

It follows that for every $h \neq g \in H$ we have

$$-\beta\rho_g\gamma_g^{-\beta+1} = -\beta\rho_h\gamma_h^{-\beta+1}, \tag{3.142}$$

and with some straightforward algebra we obtain

$$\gamma_h = \left(\frac{\rho_h}{\rho_g}\right)^{\frac{1}{\beta+1}} \gamma_g. \tag{3.143}$$

Plugging this into the equality constraint we obtain

$$\sum_{h\in H} \gamma_h = \sum_{h\in H} \left(\frac{\rho_h}{\rho_g}\right)^{\frac{1}{\beta+1}} \gamma_g = 1. \tag{3.144}$$

And solving for γ_g yields

$$\gamma_g = \frac{\rho_g^{\frac{1}{\beta+1}}}{\sum_{h\in H} \rho_h^{\frac{1}{\beta+1}}}. \tag{3.145}$$

Plugging the solution for γ into the sellers' intervention participation threshold τ, with some straightforward algebra, we obtain that the maximum τ can be over all choices of γ for a fixed set of reserve prices is

$$\tau = \left(\frac{\sum_{h\in H} \rho_h}{\sum_{h\in H} \rho_h^{\frac{1}{\beta+1}}}\right)^{\frac{\beta+1}{\beta}} c_{\mathcal{G}}. \tag{3.146}$$

This completes the first major step.

We move on to the second step. What is the maximum sellers' intervention participaiton threshold, τ, over all the possible reserve prices, q, the marketplace can set? We first derive a non-standard program and then formulate an equivalent program in standard form. First, what are the possible reserve prices the marketplace can set? In principle, the marketplace has the flexibility to set p_g to any non-negative value for each group $g \in G$, i.e., $p_g \geq 0$. Moreover p_g enters τ via $\rho_g = p_g \mathbf{1}[b_{1,g} \geq p_g]$. At intervention equilibrium, the buyer will bid its value-of-accuracy for each group $g \in G$, $\mu_{1,g} = c_\mu$. Therefore, the marketplace can set ρ_g to any value in $[0, c_\mu]$ by

setting p_g appopriately. Thus, choosing p_g is equivalent to choosing ρ_g, so we now focus on the marketplace's choices of ρ.

In this problem, ρ determines γ. γ is set to maximize the sellers' intervention participation threshold. As we have just shown, this entails that any unmonetized group has no representation, i.e., for any $g \in G \setminus H$, $\rho_g = 0$. Thus the search space over ρ is 0 if $ginG \setminus H$ and $[0, c_\mu$ if $g \in H$. We have derived the following program:

$$\max_{\rho} \left(\frac{\sum_{h \in H} \rho_h}{\sum_{h \in H} \rho_h^{\frac{1}{\beta+1}}} \right)^{\frac{\beta+1}{\beta}}, \tag{3.147}$$

subject to

$$0 \leq \rho_g \leq c_\mu, \tag{3.148}$$

for all $g \in H$, where

$$\rho_g = 0, \tag{3.149}$$

for all $g \in G \setminus H$.

We now formulate an equivalent program in standard form. Define objective,

$$f(\rho) \triangleq -\frac{\sum_{h \in H} \rho_h}{\sum_{h \in H} \rho_h^{\frac{1}{\beta+1}}}. \tag{3.150}$$

Define equality constraints: For each $g \in G \setminus H$,

$$h_g(\rho) = \rho_g, \tag{3.151}$$

subject to,

$$h_g(\rho) = 0. \tag{3.152}$$

Define inequality constraints: For each $g \in H$,

$$b_{(\ell,g)}(\rho) = -\rho_g \tag{3.153}$$

and

$$b_{(u,g)}(\rho) = \rho_g - c_\mu. \tag{3.154}$$

And define the program

$$\min_{\rho} f(\rho), \tag{3.155}$$

subject to

$$b(\rho) \leq 0. \tag{3.156}$$

We now solve the program. For each $g \in G \setminus H$, the equality constraint $h_g(\rho)$ determines $\rho_g = 0$. We turn to solving ρ_g for $g \in H$. Compute the partial derivatives of the objective and inequality constraint functions. For the objective,

$$\frac{\partial}{\partial \rho_g} f(\rho) = -\frac{\left(\sum_{h \in H} \rho_g^{\frac{1}{\beta+1}}\right) - \left(\sum_{h \in H} \rho_h\right) \frac{1}{\beta+1} \rho_g^{-\frac{\beta}{\beta+1}}}{\left(\sum_{h \in H} \rho_h^{\frac{1}{\beta+1}}\right)^2} \tag{3.157}$$

For the lower bound constraints,

$$\frac{\partial}{\partial \rho_h} b_{(\ell,g)}(\rho) = \begin{cases} -1 & \text{if } h = g \\ 0 & \text{otherwise} \end{cases} \tag{3.158}$$

And for the upper bound constraints,

$$\frac{\partial}{\partial \rho_h} b_{(u,g)}(\rho) = \begin{cases} 1 & \text{if } h = g \\ 0 & \text{otherwise} \end{cases} \tag{3.159}$$

We first find the feasible solutions that satisfy the KKT conditions. The multiplier rule of the KKT conditions gives

$$\frac{\partial}{\partial \rho_g} f(\rho) + \frac{\partial}{\partial \rho_g} b_{(\ell,g)}(\rho)\lambda_{(\ell,g)} + \frac{\partial}{\partial \rho_g} b_{(u,g)}(\rho)\lambda_{(u,g)} = 0. \tag{3.160}$$

Substituting the partial derivatives of the inequality constraints yields,

$$\frac{\partial}{\partial \rho_g} f(\rho) - \lambda_{(\ell,g)} + \lambda_{(u,g)} = 0. \tag{3.161}$$

Observe that $\rho = 0$ is not a local minimum. And for every feasible solution $\rho \neq 0$ we have

$$\frac{\partial}{\partial \rho_g} f(\rho) \neq 0. \tag{3.162}$$

It follows that if a feasible solution ρ satisfies the KKT conditions, then for all $g \in H$ either $\rho_g = 0$ or $\rho_g = c_\mu$. Let us call such a feasible solution a corner. Moreover, every corner satisfies the KKT conditions by setting their multipliers appropriately. It follows that if there is a local minimum, it is one of these corners.

We now show that every corner ρ is a local minimum. It suffices to show that for every $g \in H$ the following two implications hold: 1) if $\rho_g = 0$, then $\frac{\partial}{\partial \rho_g} f(\rho) > 0$; and 2) if $\rho_g = c_\mu$, then $\frac{\partial}{\partial \rho_g} f(\rho) < 0$. And observe that in both cases, we have

$$\frac{\partial}{\partial \rho_g} f(\rho) = -\frac{c_\mu^{\frac{1}{\beta+1}} - \frac{1}{\beta+1} c_\mu \rho_g^{-\frac{\beta}{\beta+1}}}{n(c_\mu^{\frac{1}{\beta+1}})^2}, \tag{3.163}$$

where n is the number of groups h such that $\rho_h = c_\mu$, because ρ is a corner.

Now consider the first implication. Suppose $\rho_g = 0$. Observe that $\rho_g^{-\frac{\beta}{\beta+1}} \to \infty$ as $\rho_g \to 0$. We conclude that $\frac{\partial}{\partial \rho_g} f(\rho) > 0$.

Now consider the second implication. Suppose $\rho_g = c_\mu$. Then,

$$\frac{\partial}{\partial \rho_g} f(\rho) = -\frac{\beta}{(\beta+1)n} < 0. \tag{3.164}$$

Therefore every corner ρ is a local minimum. Actually, every corner is a global minimum because since they all achieve the same objective value,

$$f(\rho) = -\frac{\sum_{h\in H} \rho_h}{\sum_{h\in H} \rho_h^{\frac{1}{\beta+1}}} = -\frac{nc_\mu}{nc_\mu^{\frac{1}{\beta+1}}} = c_\mu^{\frac{\beta}{\beta+1}}, \tag{3.165}$$

where n is the number of groups h such that $\rho_h = c_\mu$.

We conclude that the maximum sellers' intervention participation threshold is achieved when the intervention stipulates uniform intervention over the monetized groups and the marketplace sets the reserve prices as high as the buyers can bear. Since the marketplace wishes to ensure that there is representation for all the groups in the data market, this implies that the only such feasible intervention is the uniform intervention, i.e., $\gamma = u$. When the intervention is u, the sellers' intervention participation threshold becomes,

$$\tau(u) = \left(\frac{\sum_{h\in H} \rho_h}{\sum_{h\in H} \rho_h^{\frac{1}{\beta+1}}} \right)^{\frac{\beta+1}{\beta}} c_{\mathcal{G}} = c_\mu c_\kappa = \kappa^T u, \tag{3.166}$$

and the market will still form in the intervention scenario. But for any other feasible $\gamma \neq u$, this implies that

$$\tau(u) = \left(\frac{\sum_{h\in H} \rho_h}{\sum_{h\in H} \rho_h^{\frac{1}{\beta+1}}} \right)^{\frac{\beta+1}{\beta}} c_{\mathcal{G}} < c_\mu c_\kappa = \kappa^T \gamma, \tag{3.167}$$

and the market will not form in the intervention scenario. This concludes the proof. □

Theorem (3.5.3) clarifies that the backfire risk is not limited to partially forming markets. Intervention in fully forming markets can also backfire. At the same time, the backfire risk behaves qualitatively differently. Unlike in partially-forming markets, the risk is not primarily driven by

production costs. Because production costs are bounded by the potential economic values in fully forming markets, the backfire risk is driven by the relationship between potential economic values and the intervention. This indicates that restricting interventions to fully-forming markets is an effective strategy only for the uniform intervention. This may be reasonable in some real-world applications, but clearly not in others. Theorems (3.5.2) and (3.5.3) underscore the question: How can the backfire risk be mitigated in fully-forming markets? We next give sufficient conditions for a fully-forming market that ensure the market will form in the intervention scenario.

Theorem 3.5.4. *Let $\mathcal{M}$ be a fully-forming data market in the baseline scenario. Define $\eta \in [0, 1]$ to be the minimum value satisfying for all $g \in G$,*

$$\kappa_g \leq \eta \tau_g. \tag{3.168}$$

Let γ be an intervention. Define $a \geq 1$,

$$\frac{1}{a} = \min_{g \in G} \gamma_g, \tag{3.169}$$

and $b \geq 1$

$$\frac{1}{b} = \max_{g \in G} \gamma_g. \tag{3.170}$$

If the marketplace chooses target vector γ and

$$\eta < \left(\frac{b}{a}\right)^{\beta+1} \frac{1}{r|G|}, \tag{3.171}$$

then $\mathcal{M}$ will form in the intervention scenario.

Proof. We will show that there must exist a subset of groups that can be profitably monetized in the intervention scenario, i.e., there exists a strategy profile and a subset of groups such that in the intervention scenario: 1) the sellers' intervention marginal production cost is at most the sellers' intervention participation threshold; and 2) the sellers produce more than the learning ante for

every group in the subset. This implies that the market will form in the intervention scenario.

We first show that the sellers' intervention marginal production cost is at most the sellers' intervention participation threshold. Since $\mathcal{M}$ is a fully-forming data market in the baseline scenario, there exists a strategy profile $\sigma = (p, \{\mu_i\}, \{x^{(j)}\})$ such that $x_g^{([M])} > 0$ for all $g \in G$, and σ is a Nash equilibrium in the baseline scenario. Consider the following strategy profile $\sigma' = (q, \{\mu_i\}, \{n^{(j)}\})$ in the intervention scenario where $q = p$ and

$$n^{(j)} = \frac{1}{M}\left(\frac{\alpha\beta}{\kappa^T\gamma}\sum_{h\in H}\rho_h\gamma_h^{-\beta}\right)^{\frac{1}{\beta}}, \tag{3.172}$$

for some subset of monetized groups $H \subseteq G$, $H \neq \emptyset$.

Observe that we can bound the sellers' intervention production costs, $\kappa^T\gamma$, from above,

$$\kappa^T\gamma = \sum_{g\in G}\kappa_g\gamma_g \leq \sum_{g\in G}\eta\tau_g\gamma_g \leq \sum_{g\in G}\eta\tau_g\frac{1}{b} = \frac{\eta}{b}\sum_{g\in G}\rho_g c_{\mathcal{G}}. \tag{3.173}$$

The sellers' intervention participation threshold, τ, depends on the subset of monetized groups, H. And observe that we can bound τ from below,

$$\tau = \frac{\left(\sum_{g\in H}\rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(\sum_{g\in H}\rho_g\gamma_g^{-\beta}\right)^{\frac{1}{\beta}}}c_{\mathcal{G}} = \left(\frac{\sum_{g\in H}\rho_g}{\sum_{g\in H}\rho_g\gamma_g^{-\beta}}\right)^{\frac{1}{\beta}}\left(\sum_{g\in H}\rho_g\right)c_{\mathcal{G}} \tag{3.174}$$

$$\geq \left(\frac{\sum_{g\in H}\rho_g}{\sum_{g\in H}\rho_g\frac{1}{a}^{-\beta}}\right)^{\frac{1}{\beta}}\sum_{g\in H}\rho_g c_{\mathcal{G}} = \frac{1}{a}\sum_{g\in H}\rho_g c_{\mathcal{G}}. \tag{3.175}$$

Except for their index sets, the bounds are structurally very similar. We now derive an inequality to bridge the difference in their index sets. Define $\overline{H} \triangleq G \setminus H$,

$$\rho_0 \triangleq \min_{h\in H}\rho_h, \tag{3.176}$$

and

$$r \triangleq \max_{f,g \in G} \frac{\rho_f}{\rho_g}, \tag{3.177}$$

and write

$$\sum_{g \in G} \rho_g = \sum_{h \in H} \rho_h + \sum_{g \in \overline{H}} \rho_g = \sum_{h \in H} \rho_h + \sum_{g \in \overline{H}} \frac{\rho_0}{\rho_0} \rho_g = \sum_{h \in H} \rho_h + \sum_{g \in \overline{H}} \frac{\rho_g}{\rho_0} \rho_0 \tag{3.178}$$

$$\leq \sum_{h \in H} \rho_h + \sum_{g \in \overline{H}} r \rho_0 = \sum_{h \in H} \rho_h + r|\overline{H}|\rho_0 \tag{3.179}$$

$$\leq \sum_{h \in H} \rho_h + r|\overline{H}| \left(\frac{1}{|H|} \sum_{h \in H} \rho_h \right) = \left(1 + r \frac{|\overline{H}|}{|H|} \right) \sum_{h \in H} \rho_h. \tag{3.180}$$

Now $1 \leq |H| \leq |G|$, and $|\overline{H}| = |G| - |H|$ so

$$\frac{|\overline{H}|}{|H|} = \frac{|G| - |H|}{|H|} = \frac{|G|}{|H|} - 1 \leq |G| - 1, \tag{3.181}$$

and therefore

$$1 + r \frac{|\overline{H}|}{|H|} \leq 1 + r(|G| - 1) = 1 + r|G| - r \leq r|G|. \tag{3.182}$$

Putting Inequalities (3.180) and (3.182) together we have,

$$\sum_{g \in G} \rho_g \leq r|G| \sum_{h \in H} \rho_h. \tag{3.183}$$

Applying this to Inequality (3.173) we obtain,

$$\kappa^T \gamma \leq \frac{\eta}{b} r|G| \sum_{h \in H} \rho_h c_{\mathcal{G}}. \tag{3.184}$$

By definition, $b \leq a$, and by assumption it follows that

$$\eta < \left(\frac{b}{a} \right)^{\beta+1} \frac{1}{r|G|} < \frac{b}{a} \frac{1}{r|G|}. \tag{3.185}$$

Applying this to Inequality (3.184) yields,

$$\kappa^T \gamma \leq \frac{1}{a} \sum_{h \in H} \rho_h c_{\mathcal{G}}, \tag{3.186}$$

which we recognize as the lower bound on τ in Inequality (3.175), i.e.,

$$\kappa^T \gamma \leq \tau. \tag{3.187}$$

Critically, the analysis of $\kappa^T \gamma$ and τ depends on the sellers producing more than the learning ante number of samples for each group in H, that is, it requires,

$$\gamma_h n^{([M])} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}}, \tag{3.188}$$

for every $h \in H$. We now show that this holds.

$$\gamma_h n^{([M])} = \gamma_h \left(\frac{\alpha\beta}{\kappa^T \gamma} \sum_{g \in H} \rho_g \gamma_g^{-\beta}\right)^{\frac{1}{\beta+1}} \geq \frac{1}{a} \left(\frac{\alpha\beta}{\kappa^T \gamma} \sum_{g \in H} \rho_g \gamma_g^{-\beta}\right)^{\frac{1}{\beta+1}} \geq \frac{1}{a} \left(\frac{\alpha\beta}{\kappa^T \gamma} \sum_{g \in H} \rho_g b^{\beta}\right)^{\frac{1}{\beta+1}} \tag{3.189}$$

$$\geq \frac{1}{a} \left(\alpha\beta \frac{b}{\eta r |G| \sum_{h \in H} \rho_h c_{\mathcal{G}}} \sum_{g \in H} \rho_g b^{\beta}\right)^{\frac{1}{\beta+1}} = \frac{b}{a} \left(\frac{\alpha\beta}{\eta r |G| c_{\mathcal{G}}}\right)^{\frac{1}{\beta+1}} \tag{3.190}$$

$$= \frac{b}{a} \left(\frac{\alpha\beta}{\eta r |G|} \frac{\alpha^{\frac{1}{\beta}} \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}}\right)^{\frac{\beta+1}{\beta}}}{Z^{\frac{\beta+1}{\beta}}}\right)^{\frac{1}{\beta+1}} \geq \frac{b}{a} \left(\frac{1}{\eta r |G|} \left(\frac{\alpha}{Z}\right)^{\frac{\beta+1}{\beta}}\right)^{\frac{1}{\beta+1}} \tag{3.191}$$

$$= \frac{b}{a} \left(\frac{1}{\eta r |G|}\right)^{\frac{1}{\beta+1}} \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}} > \left(\frac{\alpha}{Z}\right)^{\frac{1}{\beta}} \tag{3.192}$$

because

$$\frac{b}{a} \left(\frac{1}{\eta r |G|}\right)^{\frac{1}{\beta+1}} > 1 \tag{3.193}$$

by assumption. We conclude that there exists a subset of groups H that can be profitably monetized.

□

3.6 Market growth can mitigate backfire risk and amortize the cost of fairness

In this section we study the relationship between market growth and the fairness intervention. We begin by focusing specifically on how market growth impacts the backfire risk.

In the previous section, we saw that the backfire risk is more manageable in fully-forming markets. Depending on market conditions, the marketplace may have more or less flexibility to select a target vector. Theorem (3.5.4) shows that this flexibility increases as the economic potentials of all the groups increase.

Our next result clarifies that the ability of market growth to mitigate the backfire risk is actually more flexible. It requires only that the economic potential of *a single group* grow sufficiently large. Therefore, market growth can mitigate the backfire risk in partially forming markets as well.

Claim 3.6.1. *Let γ be the marketplace's target vector. If there exists $g \in G$ such that $\max_{p_g} \rho_g \to \infty$ as $N \to \infty$, then there exists an N_0 such that $N > N_0$ implies that for all j, $\|y^{(j)}\| > 0$.*

Proof. Let $\lceil \gamma \rceil \triangleq \max_{g \in G} \gamma_g^{-\beta}$. In the intervention scenario, By Claim (3.4.2), the sellers will produce data if

$$\kappa^T \gamma \leq \frac{\left(\sum_{g \in G} \rho_g\right)^{\frac{\beta+1}{\beta}}}{\left(\sum_{g \in G} \rho_g \gamma_g^{-\beta}\right)^{\frac{1}{\beta}}} \cdot \frac{Z^{\frac{\beta+1}{\beta}}}{\alpha^{\frac{1}{\beta}} \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}}\right)^{\frac{\beta+1}{\beta}}}, \tag{3.194}$$

or equivalently

$$\frac{\alpha \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}}\right)^{\beta+1}}{Z^{\beta+1}} \left(\kappa^T \gamma\right)^{\beta} \leq \frac{\left(\sum_{g \in G} \rho_g\right)^{\beta+1}}{\left(\sum_{g \in G} \rho_g \gamma_g^{-\beta}\right)} \tag{3.195}$$

Recall that the γ_g are fixed with respect to increasing N. Bound the right-hand side of Inequality (3.195) from below by:

$$\frac{1}{\lceil \gamma \rceil} \left(\sum_{g \in G} \rho_g\right)^{\beta} \leq \frac{\left(\sum_{g \in G} \rho_g\right)^{\beta+1}}{\left(\sum_{g \in G} \rho_g \gamma_g^{-\beta}\right)} \tag{3.196}$$

Therefore, Inequality (3.195) will hold if

$$\frac{\alpha\left(\beta^{-\frac{\beta}{\beta+1}}+\beta^{\frac{1}{\beta+1}}\right)^{\beta+1}}{Z^{\beta+1}}\left(\kappa^{T}\gamma\right)^{\beta} \leq \left(\frac{1}{\lceil\gamma\rceil}\sum_{g\in G}\rho_{g}\right)^{\beta} \tag{3.197}$$

Recall that Z, α, β, and κ_g are fixed with respect to increasing N. By assumption, there is at least one group g such that $\max_{p_g} \rho_g \to \infty$ as $N \to \infty$. Observe that $\max_{p_g} \rho_g$ is non-decreasing in the number of buyers N. Therefore there is some N_0 such that Inequality (3.197) will be satisfied for all $N > N_0$. □

Taken together, Theorem (3.5.1) and Claim (3.6.1) indicate that the ability of the market to bear a fairness intervention and still function varies, and that an effective and efficient intervention must be based on an analysis of market conditions, including the potential economic value.

Market growth can mitigate the backfire risk. But after the risk is successfully mitigated, there will still be a cost of fairness, therefore it is natural to ask: What happens to the cost of fairness as the potential economic value of the market continues to grow?

Perhaps surprisingly, we find that as an infinite number of buyers enter the data market, if the potential economic value of at least one group grows unbounded, then *every agent in the data market is asymptotically at least as well off* in the intervention scenario than in the baseline scenario, and sometimes some of the buyers can be strictly better off. Stated another way: asymptotically speaking, fairness either comes at no cost or it has a positive externality.

This is driven by two inter-dependent dynamics. The first dynamic is that the growing potential economic value will eventually induce the sellers to participate in the data market. In the baseline scenario the sellers will produce training samples for any group g whose potential economic value becomes large enough; in the intervention scenario the sellers will produce training samples for every group provided that at least one group has a sufficiently large potential economic value. Moreover, the sellers will asymptotically produce an unbounded number of training samples. In the baseline scenario this only occurs for those groups whose potential economic value is asymptotically unbounded; in the intervention scenario this occurs for every group.

Asymptotically unbounded data production drives the amortization of the cost of fairness for the buyers. The sellers asymptotically produce an infinite amount of data in the intervention scenario and this has one of two consequences for each buyer's utility, depending on whether or not the sellers produce an unbounded amount of data for every group in the baseline scenario. If the sellers do, then the buyer obtains the same utility and if the sellers don't, then the buyer's utility in the intervention scenario is strictly greater than in the baseline scenario.

The second dynamic is that the sellers production costs grow sublinearly in their shares of the revenues. This is due to two factors: the free replicability of the data and the diminishing gains to accuracy as more and more samples are produced. A sample incurs a one-time fixed cost to be produced, after that it can be re-used without further cost across any number of prediction tasks. After the first sample, the second sample will increase accuracy less than the first. Hence, the sellers will slow down the rate at which they increase data production, although data production will still grow unboundedly. It turns out that this drives the amortization of the cost of fairness for the sellers and the marketplace.

We now present the formal analysis, focusing first on the dynamic of asympotically unbounded data production. Claim (3.6.1) gives a condition under which the sellers will eventually participate in the market in the intervention scenario. Under the same condition, the sellers will eventually participate in the market in the baseline scenario.

Claim 3.6.2. *If* $\max_{p_g} \rho_g \to \infty$ *as* $N \to \infty$*, then there exists an* N_0 *such that* $N > N_0$ *implies that for all* j, $x_g^{(j)} > 0$.

Proof. In the baseline scenario, the sellers will produce data if

$$\kappa_g \leq \tau_g = \frac{\rho_g Z^{\frac{\beta+1}{\beta}}}{\alpha^{\frac{1}{\beta}} \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}}\right)^{\frac{\beta+1}{\beta}}}, \tag{3.198}$$

or equivalently

$$\frac{\alpha^{\frac{1}{\beta}} \left(\beta^{-\frac{\beta}{\beta+1}} + \beta^{\frac{1}{\beta+1}}\right)^{\frac{\beta+1}{\beta}}}{Z^{\frac{\beta+1}{\beta}}} \kappa_g \leq \rho_g \tag{3.199}$$

Recall that Z, α, β, and κ_g are fixed with respect to increasing N. By assumption, $\max_{p_g} \rho_g \to \infty$ as $N \to \infty$ and note that $\max_{p_g} \rho_g$ is non-decreasing in the number of buyers N, therefore there is some N_0 such that Inequality (3.199) is satisfied for every $N > N_0$. □

Although Claims (3.6.1) and (3.6.2) indicate that the same condition asymptotically induces seller participation in both scenarios, it is important to note that it indicates the induced participation is uneven: in the baseline scenario the sellers are only induced to produce data for those groups whose potential economic value becomes sufficiently large; whereas in the intervention scenario the sellers are induced to produce data for every group, provided that the potential economic value of at least one group becomes sufficiently large. Claims (3.6.1) and (3.6.2) indicate that the sellers will eventually produce some training samples, but they afford little indication as to how many. For that, we have the following result in the baseline scenario.

Claim 3.6.3. *If* $\max_{p_g} \rho_g \to \infty$ *as* $N \to \infty$, *then* $x_g^{([M])} \to \infty$.

Proof. By Lemma (3.4.1), at equilibrium we have that

$$x_g^{([M])} = \left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{1/(\beta+1)}. \tag{3.200}$$

By assumption, $\max_{p_g} \rho_g \to \infty$, it follows that by Claim (3.6.2) the sellers will participate in the market for all N sufficiently large and therefore

$$x_g^{([M])} = \left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{1/(\beta+1)} \to \infty. \tag{3.201}$$

□

And we have a corresponding result in the intervention scenario.

Claim 3.6.4. *If there exists* $g \in G$ *such that* $\max_{p_g} \rho_g \to \infty$ *as* $N \to \infty$, *then* $n^{([M])} = \|y^{([M])}\| \to \infty$.

Proof. By Lemma (3.4.2) and Claim (3.4.2) the sellers will produce

$$n^{([M])} = \left(\frac{\alpha\beta}{\kappa^T \gamma} \sum_{g \in H} \rho_g \gamma_g^{-\beta} \right)^{1/(\beta+1)} \tag{3.202}$$

samples at equilibrium if

$$\kappa^T \gamma \leq \tau_H(\rho, \gamma) \tag{3.203}$$

where

$$H \triangleq \{g \in G : \gamma_g \geq \gamma_h\}, \tag{3.204}$$

and γ_h is a minimum value over γ_g satisfying

$$\gamma_g n^{([M])} > \left(\frac{\alpha}{Z} \right)^{\frac{1}{\beta}}. \tag{3.205}$$

Observe that $\kappa^T \gamma$ is fixed as $\rho_g \to \infty$, whereas $n^{([M])}$ and $\tau_G(\rho, \gamma)$ grow unboundedly. □

Claims (3.6.3) and (3.6.4) indicate that if the potential economic value of some group grows asymptotically unbounded, then data production also grows asymptotically unbounded. Moreover, eventually the marketplace's strategy will be the same in both scenarios. This drives the amortization of the cost of fairness to the buyers.

Theorem 3.6.1. *If there exists $g \in G$ such that $\max_{p_g} \rho_g \to \infty$ as $N \to \infty$, then for every buyer i we have*

$$\lim_{N \to \infty} \frac{u_i^f(\mu_i)}{u_i(\mu_i)} \geq 1. \tag{3.206}$$

Proof. Consider buyer i's utility in the baseline scenario,

$$u_i(\mu_i) = \sum_{g \in G} (\mu_{i,g} - p_g) \mathcal{G}(\mathcal{AF}_g(\mu_{i,g}, x^{([M])})). \tag{3.207}$$

Buyer i will be allocated data for group g if and only if $\mu_{i,g} \geq p_g$. Let $G_p \triangleq \{g \in G : \mu_{i,g} \geq p_g\}$,

and consider the buyer's utility in the limit,

$$\lim_{N\to\infty} u_i(\mu_i) = \lim_{N\to\infty} \sum_{g\in G_p} (\mu_{i,g} - p_g)\mathcal{G}(x_g^{([M])}) \leq Z \sum_{g\in G_p} (\mu_{i,g} - p_g), \tag{3.208}$$

where the inequality follows because for all g and $x_g^{([M])}$, $\mathcal{G}(x_g^{([M])}) \leq Z$. By assumption, there is at least one group such that $\max_{p_g} \rho_g \to \infty$ as $N \to \infty$. For any such group it follows that $\mathcal{G}(x_g^{([M])}) \to Z$ as $N \to \infty$ because $x_g^{([M])} \to \infty$ by Claim (3.6.3).

Now consider buyer i's utility in the intervention scenario.

$$u_i^f(\mu_i) = \sum_{g\in G} (\mu_{i,g} - p_g)\mathcal{G}(\mathcal{AF}_g(\mu_{i,g}, y^{([M])})). \tag{3.209}$$

Again, buyer i will be allocated data for group g if and only if $\mu_{i,g} \geq p_g$. Therefore, the buyer's utility in the limit is

$$\lim_{N\to\infty} u_i^f(\mu_i) = \lim_{N\to\infty} \sum_{g\in G_p} (\mu_{i,g} - p_g)\mathcal{G}(y_g^{([M])}) = Z \sum_{g\in G_p} (\mu_{i,g} - p_g). \tag{3.210}$$

The last equality follows from the assumption that there is at least one group such that $\max_{p_g} \rho_g \to \infty$ as $N \to \infty$. By Claim (3.6.4), it follows that for every group g, $y_g^{([M])} \to \infty$ and consequently $\mathcal{G}(y_g^{([M])}) \to Z$.

Putting these together we can evaluate the cost of fairness in the limit

$$\lim_{N\to\infty} \frac{u_i^f(\mu_i)}{u_i(\mu_i)} \geq \frac{Z \sum_{g\in G_p}(\mu_{i,g} - p_g)}{Z \sum_{g\in G_p}(\mu_{i,g} - p_g)} = 1. \tag{3.211}$$

□

It is sufficient for the economic potential of only one group to grow asymptotically unbounded to either amortize the cost of fairness to the buyers or make the buyers better off in the intervention scenario asymptotically. This condition is also sufficient to amortize the cost of fairness to the sellers and the marketplace. But the amortization is now driven by the second dynamic, slow

growth in production costs, interacting with the first dynamic of unbounded data production. We first analyze the case of the sellers.

Theorem 3.6.2. *If there exists $g \in G$ such that $\max_{p_g} \rho_g \to \infty$ as $N \to \infty$, then for every seller j we have*

$$\lim_{N\to\infty} \frac{v_j^f(y^{(j)})}{v_j(x^{(j)})} = 1. \tag{3.212}$$

Proof. Consider seller j's utility in the baseline scenario.

$$v_j(x^{(j)}) = \sum_{g\in G} p_g \sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \sum_{T\subseteq[M]\setminus\{j\}} c_T \left(\mathcal{G}(x_g^{(T\cup\{j\})}) - \mathcal{G}(x_g^{(T)})\right) - \sum_{g\in G} \kappa_g x_g^{(j)} \tag{3.213}$$

$$= \sum_{g\in G} \rho_g \frac{1}{M} \mathcal{G}(x_g^{([M])}) - \sum_{g\in G} \kappa_g x_g^{(j)} \tag{3.214}$$

$$= \sum_{g\in G} \rho_g \frac{1}{M} \left(Z - \alpha (x_g^{([M])})^{-\beta}\right) - \sum_{g\in G} \kappa_g x_g^{(j)} \tag{3.215}$$

$$= \frac{Z}{M} \sum_{g\in G} \rho_g - \frac{\alpha}{M} \sum_{g\in G} \rho_g \left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{-\frac{\beta}{\beta+1}} - \sum_{g\in G} \kappa_g \frac{1}{M} \left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}} \tag{3.216}$$

$$= \frac{Z}{M} \sum_{g\in G} \rho_g - \frac{\alpha}{M} \sum_{g\in G} \rho_g^{\frac{1}{\beta+1}} \left(\frac{\alpha\beta}{\kappa_g}\right)^{-\frac{\beta}{\beta+1}} - \sum_{g\in G} \kappa_g \frac{1}{M} \left(\frac{\rho_g}{\kappa_g}\alpha\beta\right)^{\frac{1}{\beta+1}} \tag{3.217}$$

Note that the positive term is linear in the sum of the ρ_g whereas the two negative terms are

sublinear. Now consider seller j's utility in the intervention scenario.

$$v_j^f(y^{(j)}) = \sum_{g\in G} p_g \sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \sum_{T\subseteq[M]\setminus\{j\}} c_T \left(\mathcal{G}(y_g^{(T\cup\{j\})}) - \mathcal{G}(y_g^{(T)})\right) - \sum_{g\in G} \kappa_g y_g^{(j)} \tag{3.218}$$

$$= \sum_{g\in G} \rho_g \frac{1}{M} \mathcal{G}(y_g^{([M])}) - \sum_{g\in G} \kappa_g y_g^{(j)} \tag{3.219}$$

$$= \sum_{g\in G} \rho_g \frac{1}{M} \left(Z - \alpha (y_g^{([M])})^{-\beta}\right) - \sum_{g\in G} \kappa_g y_g^{(j)} \tag{3.220}$$

$$= \frac{Z}{M} \sum_{g\in G} \rho_g - \frac{\alpha}{M} \sum_{g\in G} \rho_g \gamma_g^{-\beta} \left(\frac{\alpha\beta}{\kappa^T\gamma} \sum_{g\in G} \rho_g \gamma_g^{-\beta}\right)^{-\frac{\beta}{\beta+1}} - \sum_{g\in G} \frac{\kappa_g \gamma_g}{M} \left(\frac{\alpha\beta}{\kappa^T\gamma} \sum_{g\in G} \rho_g \gamma_g^{-\beta}\right)^{\frac{1}{\beta+1}} \tag{3.221}$$

$$= \frac{Z}{M} \sum_{g\in G} \rho_g - \frac{\alpha}{M} \frac{\alpha\beta}{\kappa^T\gamma}^{-\frac{\beta}{\beta+1}} \left(\sum_{g\in G} \rho_g \gamma_g^{-\beta}\right)^{\frac{1}{\beta+1}} - \sum_{g\in G} \frac{\kappa_g \gamma_g}{M} \left(\frac{\alpha\beta}{\kappa^T\gamma} \sum_{g\in G} \rho_g \gamma_g^{-\beta}\right)^{\frac{1}{\beta+1}} \tag{3.222}$$

Note that the positive term is linear in the sum of the ρ_g where as the two negative terms are sublinear. Consequently, seller j's utility ratio in the limit is

$$\lim_{N\to\infty} \frac{v_j^f(y^{(j)})}{v_j(x^{(j)})} = \frac{\frac{Z}{M}\sum_{g\in G}\rho_g}{\frac{Z}{M}\sum_{g\in G}\rho_g} = 1. \tag{3.223}$$

□

And now for the marketplace.

Theorem 3.6.3. *If there exists $g \in G$ such that $\max_{p_g} \rho_g \to \infty$ as $N \to \infty$, then for the marketplace we have*

$$\lim_{N\to\infty} \frac{w^f(p)}{w(p)} = 1 \tag{3.224}$$

Proof. Consider the marketplace's utility in the baseline scenario.

$$w(p) = \sum_{g \in G} p_g \sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \mathcal{G}(x_g^{([M])}) \tag{3.225}$$

$$= \sum_{g \in G} \rho_g \left(Z - \alpha (x_g^{([M])})^{-\beta} \right) \tag{3.226}$$

$$= Z \sum_{g \in G} \rho_g - \alpha \sum_{g \in G} \rho_g (x_g^{([M])})^{-\beta} \tag{3.227}$$

$$= Z \sum_{g \in G} \rho_g - \alpha \sum_{g \in G} \rho_g \left(\frac{\rho_g}{\kappa_g} \alpha \beta \right)^{-\frac{\beta}{\beta+1}} \tag{3.228}$$

$$= Z \sum_{g \in G} \rho_g - \alpha \sum_{g \in G} \rho_g^{\frac{1}{\beta+1}} \left(\frac{\alpha \beta}{\kappa_g} \right)^{-\frac{\beta}{\beta+1}} \tag{3.229}$$

Note that the positive term is linear in the sum of the ρ_g whereas the negative term is sublinear. Now consider the marketplace's utility in the intervention scenario.

$$w^f(p) = \sum_{g \in G} p_g \sum_{i=1}^{N} \mathbf{1}[\mu_{i,g} \geq p_g] \mathcal{G}(y_g^{([M])}) \tag{3.230}$$

$$= \sum_{g \in G} \rho_g \left(Z - \alpha (y_g^{([M])})^{-\beta} \right) \tag{3.231}$$

$$= Z \sum_{g \in G} \rho_g - \alpha \sum_{g \in G} \rho_g (y_g^{([M])})^{-\beta} \tag{3.232}$$

$$= Z \sum_{g \in G} \rho_g - \alpha \sum_{g \in G} \rho_g \gamma_g^{-\beta} \left(\frac{\alpha \beta}{\kappa^T \gamma} \sum_{g \in G} \rho_g \gamma_g^{-\beta} \right)^{-\frac{\beta}{\beta+1}} \tag{3.233}$$

$$= Z \sum_{g \in G} \rho_g - \frac{\alpha}{M} \frac{\alpha \beta}{\kappa^T \gamma}^{-\frac{\beta}{\beta+1}} \left(\sum_{g \in G} \rho_g \gamma_g^{-\beta} \right)^{\frac{1}{\beta+1}} \tag{3.234}$$

Note that the positive term is linear in the sum of the ρ_g whereas the negative term is sublinear. Consequently, the marketplace's utility ratio in the limit is

$$\lim_{N \to \infty} \frac{w^f(p)}{w(p)} = \lim_{N \to \infty} \frac{Z \sum_{g \in G} \rho_g}{Z \sum_{g \in G} \rho_g} = 1 \tag{3.235}$$

□

References

[1] M. Chui, E. Hazan, R. Roberts, A. Singla, and K. Smaje, "The economic potential of generative ai," 2023.

[2] T. Härlin, G. B. Rova, A. Singla, O. Sokolov, and A. Sukharevsky, "Exploring opportunities in the generative ai value chain," 2023.

[3] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE intelligent systems*, vol. 24, no. 2, pp. 8–12, 2009.

[4] A. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, 2001.

[5] A. Gelman, J. Hill, and A. Vehtari, *Regression and other stories*. Cambridge University Press, 2021.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[7] T. Viering and M. Loog, "The shape of learning curves: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7799–7819, 2022.

[8] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[9] J. Hoffmann *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[10] D. Seetharaman, "For data-guzzling ai companies, the internet is too small," *The Wall Street Journal*, Apr. 1, 2024.

[11] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[12] G. Team *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[13] C. Metz, C. Kang, S. Frenkel, S. A. Thompson, and N. Grant, "How tech giants cut corners to harvest data for a.i.," *The New York Times*, Apr. 6, 2024.

[14] K. Paul and A. Tong, "Inside big tech's underground race to buy ai training data," *Reuters*, Apr. 5, 2024.

[15] P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho, "Will we run out of data? an analysis of the limits of scaling datasets in machine learning," *arXiv preprint arXiv:2211.04325*, 2022.

[16] K. Paul, "Exclusive: Multiple ai companies bypassing web standard to scrape publisher sites, licensing firm says," *Reuters*, Jun. 21, 2024.

[17] S. Fischer, "Major websites are blocking ai crawlers from accessing their content," *Axios*, Aug. 31, 2023.

[18] M. M. Grynbaum and R. Mac, "The times sues openai and microsoft over a.i. use of copyrighted work," *The New York Times*, Dec. 27, 2023.

[19] M. O'Brien, "Chatgpt-maker openai signs deal with ap to license news stories," *The Associated Press*, Jul. 13, 2023.

[20] S. A. Azcoitia and N. Laoutaris, "A survey of data marketplaces and their business models," *ACM SIGMOD Record*, vol. 51, no. 3, pp. 18–29, 2022.

[21] F. Schomm, F. Stahl, and G. Vossen, "Marketplaces for data: An initial survey," *ACM SIGMOD Record*, vol. 42, no. 1, pp. 15–26, 2013.

[22] J. Kennedy, P. Subramaniam, S. Galhotra, and R. Castro Fernandez, "Revisiting online data markets in 2022: A seller and buyer perspective," *ACM SIGMOD Record*, vol. 51, no. 3, pp. 30–37, 2022.

[23] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "An online pricing mechanism for mobile crowdsensing data markets," in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2017, pp. 1–10.

[24] M. Zhang, F. Beltrán, and J. Liu, "A survey of data pricing for data marketplaces," *IEEE Transactions on Big Data*, 2023.

[25] D. Bergemann and A. Bonatti, "Markets for information: An introduction," *Annual Review of Economics*, vol. 11, pp. 85–107, 2019.

[26] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019, pp. 701–726.

[27] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography: Third Theory of Cryptography Conference,*

TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3, Springer, 2006, pp. 265–284.

[28] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proceedings of the 12th ACM conference on Electronic commerce*, 2011, pp. 199–208.

[29] 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*, May 23, 2016.

[30] M. Varner and A. Sankin, "Suckers list: How allstate's secret auto insurance algorithm squeezes big spenders," *The Markup*, Feb. 25, 2020.

[31] A. Tobin, "Hud sues facebook over housing discrimination and says the company's algorithms have made the problem worse," *ProPublica*, Mar. 28, 2019.

[32] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[33] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[34] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[35] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[36] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[37] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, accountability and transparency*, PMLR, 2018, pp. 107–118.

[38] B. d'Alessandro, C. O'Neil, and T. LaGatta, "Conscientious classification: A data scientist's guide to discrimination-aware classification," *Big data*, vol. 5, no. 2, pp. 120–134, 2017.

[39] E. Black, "(un) fairness along the ai pipeline problems and solutions," Ph.D. dissertation, Carnegie Mellon University.

[40] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*, PMLR, 2013, pp. 325–333.

[41] R. H. Coase, *The nature of the firm*. Springer, 1995.

[42] L. Gelauff, A. Goel, K. Munagala, and S. Yandamuri, "Advertising for demographically fair outcomes," *arXiv preprint arXiv:2006.03983*, 2020.

[43] A. Lambrecht and C. Tucker, "Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads," *Management science*, vol. 65, no. 7, pp. 2966–2981, 2019.

[44] J. Angwin, A. Tobin, and M. Varner, "Facebook (Still) Letting Housing Advertisers Exclude Users by Race," *ProPublica*, Nov. 21, 2017.

[45] A. Datta, A. Datta, J. Makagon, D. K. Mulligan, and M. C. Tschantz, "Discrimination in online advertising: A multidisciplinary inquiry," in *Conference on Fairness, Accountability and Transparency*, PMLR, 2018, pp. 20–34.

[46] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," *arXiv preprint arXiv:1408.6491*, 2014.

[47] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Guilt-free data reuse," *Communications of the ACM*, vol. 60, no. 4, pp. 86–93, 2017.

[48] A Tobin and J. Merrill, *Facebook is letting job advertisers target only men–propublica*, 2018.

[49] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *International Conference on Machine Learning*, PMLR, 2018, pp. 3384–3393.

[50] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[51] J. E. Johndrow and K. Lum, "An algorithm for removing sensitive information," *The Annals of Applied Statistics*, vol. 13, no. 1, pp. 189–220, 2019.

[52] H. Edwards and A. Storkey, "Censoring representations with an adversary," *arXiv preprint arXiv:1511.05897*, 2015.

[53] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," *arXiv preprint arXiv:1707.00075*, 2017.

[54] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[55] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.

[56] D. McNamara, C. S. Ong, and R. C. Williamson, "Provably fair representations," *arXiv preprint arXiv:1710.04394*, 2017.

[57] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, "The measure and mismeasure of fairness," *The Journal of Machine Learning Research*, vol. 24, no. 1, pp. 14 730–14 846, 2023.

[58] OSTP. "Blueprint for an ai bill of rights." (2022), (visited on 02/06/2024).

[59] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd international conference on computer, control and communication*, IEEE, 2009, pp. 1–6.

[60] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*, PMLR, 2017, pp. 962–970.

[61] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," *Advances in neural information processing systems*, vol. 30, 2017.

[62] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data mining and knowledge discovery*, vol. 21, pp. 277–292, 2010.

[63] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *2010 IEEE international conference on data mining*, IEEE, 2010, pp. 869–874.

[64] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" *Advances in neural information processing systems*, vol. 31, 2018.

[65] S. Chawla and M. Jagadeesan, "Individual fairness in advertising auctions through inverse proportionality," in *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

[66] C. Ilvento, M. Jagadeesan, and S. Chawla, "Multi-category fairness in sponsored search auctions," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 348–358.

[67] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[68] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, PMLR, 2018, pp. 77–91.

[69] H. Elzayn and B. Fish, "The effects of competition and regulation on error inequality in data-driven markets," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 669–679.

[70] F. T. Commission *et al.*, "Data brokers: A call for transparency and accountability," *Washington, DC*, vol. 20, 2014.

[71] W. Cai *et al.*, "Adaptive sampling strategies to construct equitable training datasets," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1467–1478.

[72] G. Liao, Y. Su, J. Ziani, A. Wierman, and J. Huang, "The privacy paradox and optimal bias-variance trade-offs in data acquisition," *ACM SIGMETRICS Performance Evaluation Review*, vol. 49, no. 2, pp. 6–8, 2022.

[73] R. Maio and A. Chaintreau, "Incentives needed for low-cost fair lateral data reuse," in *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, 2020, pp. 71–82.

[74] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[75] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke, "Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–30, 2019.

[76] L. Sweeney, "Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising," *Queue*, vol. 11, no. 3, pp. 10–29, 2013.

[77] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," in *International Conference on Machine Learning*, PMLR, 2018, pp. 3150–3158.

[78] Amazon. "Aws data exchange." (), (visited on 08/08/2024).

[79] Snowflake. "Snowflake marketplace." (), (visited on 08/08/2024).

[80] Dawex. "Dawex data marketplace." (), (visited on 08/08/2024).

[81] Amazon. "Amazon bedrock." (), (visited on 08/08/2024).

[82] Google. “Gemini.” (), (visited on 08/08/2024).

[83] S. AI. “Stability ai.” (), (visited on 08/08/2024).

[84] Jammable. “Jammable.” (), (visited on 08/08/2024).

[85] OpenAI. “Chatgpt.” (), (visited on 08/08/2024).

[86] U. C. Bureau. “What we do.” (), (visited on 08/08/2024).

[87] U. C. Bureau. “Our surveys & programs.” (), (visited on 08/08/2024).

[88] U. C. Bureau. “Data stewardship.” (), (visited on 08/08/2024).

[89] U. C. Bureau. “Statistical safeguards.” (), (visited on 08/08/2024).

[90] U. C. Bureau. “Census datasets.” (), (visited on 08/12/2024).

[91] Techmap. “Job posting datasets for the united states (us).” (), (visited on 08/12/2024).

[92] Mozilla. “Datasets.” (), (visited on 08/12/2024).

[93] V. Pratap *et al.*, “Scaling speech technology to 1,000+ languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[94] Mozilla. “Common voice.” (), (visited on 08/15/2024).

[95] U. C. Bureau, *Fiscal year 2025 budget summary*, U.S. Census Bureau, Washington, DC, USA, 2024.

[96] LiveRamp, *2024 annual report*, LiveRamp, San Francisco, CA, USA, 2024.

[97] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, “Decoupled classifiers for fair and efficient machine learning,” *arXiv preprint arXiv:1707.06613*, 2017.

[98] K. J. Arrow, *Economic welfare and the allocation of resources for invention*. Springer, 1972.

[99] R. Castro Fernandez, “Protecting data markets from strategic buyers,” in *Proceedings of the 2022 International Conference on Management of Data*, 2022, pp. 1755–1769.

[100] R. B. Myerson, “Optimal auction design,” *Mathematics of operations research*, vol. 6, no. 1, pp. 58–73, 1981.

[101] N. H. Abel, "Démonstration de l'impossibilité de la résolution algébrique des équations générales qui passent le quatrieme degré," *Journal für die reine und angewandte Mathematik*, vol. 1, pp. 65–96, 1826.

[102] B. Van Der Waerden and F. Blum, *Modern algebra. Volume I.* Frederick Ungar, 1949.

[103] D. Berend and A. Kontorovich, "A sharp estimate of the binomial mean absolute deviation with applications," *Statistics & Probability Letters*, vol. 83, no. 4, pp. 1254–1259, 2013.

[104] A. W. Van Der Vaart and J. A. Wellner, *Weak convergence*. Springer, 1996.

Appendix A: Probabilistic Inequalities

Theorem A.0.1. *(Berend-Kontorovich Inequality [103]) Let X be a binomial random variable with parameters n and p such that* $1/n \leq p \leq 1 - (1/n)$*, then*

$$\sqrt{\frac{Var\left[X\right]}{2}} \leq \mathbb{E}\left[|X - \mathbb{E}\left[X\right]|\right] \leq \sqrt{Var\left[X\right]}. \tag{A.1}$$

Theorem A.0.2. *(Bretagnolle-Huber-Carol Inequality [104]) Let X be a k-dimensional multinomial random vector with parameters n and* $p \in \mathbb{R}^k$*, then*

$$\Pr\left[\sum_{i=1}^{k} |X_i - np_i| \geq 2\sqrt{n}\lambda\right] \leq 2^k \exp(-2\lambda^2), \tag{A.2}$$

$\lambda > 0$.