Incentives Needed for Low-Cost Fair Lateral Data Reuse

Roland Maio*
Columbia University in the City of New York
New York, New York
roland@cs.columbia.edu

Augustin Chaintreau
Columbia University in the City of New York
New York, New York
augustin@cs.columbia.edu

ABSTRACT

A central goal of algorithmic fairness is to build systems with fairness properties that compose gracefully. A major effort and step towards this goal in data science has been the development of *fair representations* which guarantee demographic parity under sequential composition by imposing a *demographic secrecy* constraint. In this work, we elucidate limitations of demographically secret fair representations and propose a fresh approach to potentially overcome them by incorporating information about parties' incentives into fairness interventions. Specifically, we show that in a stylized model, it is possible to relax demographic secrecy to obtain *incentive-compatible representations*, where rational parties obtain exponentially greater utilities vis-à-vis any demographically secret representation and satisfy demographic parity. These substantial gains are recovered not from the well-known *cost of fairness*, but rather from a *cost of demographic secrecy* which we formalize and quantify for the first time. We further show that the sequential composition property of demographically secret representations is not robust to aggregation. Our results open several new directions for research in fair composition, fair machine learning and algorithmic fairness.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Learning latent representations*; • **Theory of computation** → *Machine learning theory*; **Theory and algorithms for application domains**; Error-correcting codes; *Generating random combinatorial structures*; • **Mathematics of computing** → Combinatorics;

KEYWORDS

algorithmic fairness, fair machine learning, incentives, fair representations, cost of fairness, cost of demographic secrecy

ACM Reference Format:
Roland Maio and Augustin Chaintreau. 2020. Incentives Needed for Low-Cost Fair Lateral Data Reuse. In *Proceedings of the 2020 ACM-IMS Foundations of Data Science Conference (FODS '20), October 19–20, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3412815.3416890

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)[6]. 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[1]. 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[12]. 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[1, 7–9, 14, 19].

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 [21]. Our choice rules out, for

[*] This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. (DGE-1644869). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

FODS '20, October 19–20, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8103-1/20/10…$15.00
https://doi.org/10.1145/3412815.3416890

[1] 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.

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 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 [21] 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 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 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 incentives 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 6)

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

2 RELATED WORK

Our work is situated in the literature on *fair representations* initiated by Zemel et al. in [21], 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 [16], 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[21]. Feldman et al. propose an approach that learns a transport map from each group's distribution to the aggregate empirical distribution of the data[11]. Johndrow and Lum generalize this to a statistical model-based approach capable of handling discrete features and an arbitrary choice of target distribution[13]. A number of papers have considered approaches based on adversarial learning with variations in the choice of generator, adversary, and their respective optimization objectives[3, 10, 15, 16]. 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[4], 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[17]. 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[5, 18]. 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.

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 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 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 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

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

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[4], 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[4], as one of the most prominent 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). We will 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 3.4. (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

$$\Pr\left[D_i(r(u)) = o | \gamma(u) = g\right] = \Pr\left[D_i(r(v)) = o | \gamma(v) = g'\right], \tag{2}$$

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:

$$u(Z) = \sum_{i=1}^{n} \max_{D \in \mathcal{D}_i^Z} u_i(D) \tag{3}$$

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

Definition 3.5. (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

$$u(Z) = \max_{Z' \in \Pi(V)} u(Z'), \tag{4}$$

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 3.6. (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

$$\frac{|z_g|}{|z|} = \frac{|V_g|}{|V|}. \tag{5}$$

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 3.7. (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:

$$\text{If } D_i \in \arg\max_{D \in \mathcal{D}_i^Z} u_i(D), \text{ then } D_i \text{ satisfies demographic parity.} \tag{6}$$

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

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 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

$$\delta(V, U) = \max_{Z \in \Xi(V)} u(Z). \tag{7}$$

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 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

$$\beta(V, U) = \sum_{i=1}^{n} \beta_i, \tag{8}$$

where

$$\beta_i = \max_{D_i \in \mathcal{D}_i^I} u_i(D_i), \tag{9}$$

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

Definition 4.3. (Cost of Fairness) Let (V, U) be an instance of the fair representation problem. The *cost of fairness* is defined to be

$$\mathrm{CoF}(V, U) = u(I) - \beta(V, U). \tag{10}$$

Definition 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

$$\mathrm{CDS}(V, U) = \beta - \delta. \tag{11}$$

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

$$\mathrm{rCDS}(V, U) = \frac{\mathrm{CDS}(V, U)}{\beta} = \frac{\beta - \delta}{\beta}, \tag{12}$$

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.

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 4.5. *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 4.6. (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 4.5 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.

4.2 Multiple Data Consuming Firms

Theorem 4.5 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 4.7. *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 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

$$\mathrm{CDS}(V, U) > 0. \tag{13}$$

□

Table 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 4.7 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 4.8. *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 4.9. *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*

$$rCDS(V, U) \geq \frac{1}{4}. \tag{14}$$

Our leading discussion captures the key insights of the proofs which we defer to the Supplementary Materials.

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 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 $\{f_i : V \mapsto \{0, 1\}\} = \mathrm{RFM}_p(V)$, 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 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*

$$\mathbb{E}\left[CoF(V, U)\right] = \Theta\left(n\sqrt{|V|}\right). \tag{15}$$

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

Theorem 5.3. *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*

$$\mathbb{E}\left[CDS(V, U)\right] = \Omega\left(2^{n/2}\sqrt{|V|}\right). \tag{16}$$

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 5.4. *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.[2] We argue that $n = O(\log|V|)$ is a sensible and interesting choice. Consider our motivating example of targeted online advertising, constant n

[2] 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.

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, 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 4.8 (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. We present the proof in the Supplementary Material.

Lemma 5.5. *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*

$$\mathbb{E}\left[|X_0 - X_1|\right] = \Theta\left(\sqrt{Var[X_0]}\right). \tag{17}$$

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

Definition 5.6. (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

$$\epsilon(f) = \sum_{u \in V_0} f(u) - \sum_{v \in V_1} f(v). \tag{18}$$

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 5.5 establishes Theorem 5.2.

We now turn to proving Theorem 5.3. 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 5.7. (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

$$c(Z) = \sum_{z \in Z} \sum_{u, v \in z} \sum_{i \in [n]} \mathbf{1}\left[f_i(u) \neq f_i(v)\right]. \tag{19}$$

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 5.8. *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,*

$$c(Z) \geq \frac{1}{2} \sum_{\ell \in L} |\epsilon(\ell)|, \tag{20}$$

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

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

Lemma 5.9. *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*

$$\mathbb{E}\left[\sum_{\ell \in L} |\epsilon(\ell)|\right] = \Theta\left(2^{n/2}\sqrt{|V|}\right), \tag{21}$$

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

$$X_{g,\ell} := \sum_{u \in V_g} \mathbf{1}\left[\ell = k(u)\right], \tag{22}$$

and

$$Y := X_0 - X_1, \tag{23}$$

so that

$$\|Y\|_1 = \sum_{\ell \in L} |\epsilon(\ell)|. \tag{24}$$

Write

$$\mathbb{E}\left[\|Y\|_1\right] = \mathbb{E}\left[\sum_{\ell \in L} |Y_\ell|\right] = 2^n \mathbb{E}\left[\left|X_{0,\ell} - X_{1,\ell}\right|\right]. \tag{25}$$

Note that $X_{0,\ell}$ is a binomial random variable with parameters $|V|/2$ and 2^{-n}, and apply Lemma 5.5 to obtain,

$$\mathbb{E}\left[\|Y\|_1\right] = 2^n \Theta\left(\sqrt{\frac{|V|}{2}\frac{1}{2^n}\left(1 - \frac{1}{2^n}\right)}\right) = \Theta\left(2^{n/2}\sqrt{|V|}\right). \tag{26}$$

□

Theorem 5.3 follows as a consequence. Finally, we turn to proving Theorem 5.4. 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 5.10. *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

$$X_{g,i,u} \coloneqq f_i(u), \tag{27}$$

and

$$X_{g,i} \coloneqq \sum_{u \in V_g} X_{g,i,u}. \tag{28}$$

Then

$$\epsilon(f_i) = X_{0,i} - X_{1,i}. \tag{29}$$

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

$$\Pr\left[\left|X_{g,i} - \mathbb{E}\left[X_{g,i}\right]\right| \geq \sqrt{|V|\ln(40n)/4}\right] \leq \frac{1}{20n}. \tag{30}$$

Define indicator random variables

$$Y_{g,i} \coloneqq \begin{cases} 1 & \left|X_{g,i} - \mathbb{E}\left[X_{g,i}\right]\right| \geq \sqrt{|V|\ln(40n)/4} \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

and

$$Y \coloneqq \sum_{g \in G} \sum_{i \in [n]} Y_{g,i}. \tag{32}$$

By Markov's Inequality we have

$$\Pr\left[Y \geq 1\right] \leq \frac{1}{10} \tag{33}$$

and therefore

$$\Pr\left[Y < 1\right] \geq \frac{9}{10}, \tag{34}$$

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

$$|\epsilon(f_i)| \leq 2\sqrt{|V|\ln(40n)/4} = \sqrt{|V|\ln(40n)}. \tag{35}$$

□

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 5.11. *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*

$$\frac{|V|^{3/4}}{10} \tag{36}$$

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 Supplementary Materials), we have

$$\Pr\left[\sum_{j=1}^{2^n}\left|X_{0,j} - \frac{|V|}{2 \cdot 2^{-n}}\right| \geq 2\sqrt{\frac{|V|}{2}}\lambda\right] \leq 2^{2^n}\exp(-2\lambda^2). \tag{37}$$

We desire

$$2^{2^n}\exp(-2\lambda^2) \leq \frac{1}{10} \tag{38}$$

Solving for a lower bound on lambda we obtain

$$\lambda \geq \frac{5 \cdot 2^n}{8 \log e}. \tag{39}$$

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

$$\sqrt{2|V|}\lambda, \tag{40}$$

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

$$\frac{|V|}{2 \cdot 2^n} - 2\sqrt{\frac{|V|}{2}}\lambda \implies \tag{41}$$

$$\frac{|V|^{3/4}}{2} - \frac{5|V|^{3/4}}{8\log e} > \frac{|V|^{3/4}}{10} \tag{42}$$

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 5.4, 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

$$\frac{|V|^{3/4}}{10} \tag{43}$$

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. □

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.

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 artefact 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.

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.

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[3].

[3] To see this, consider the example given in Theorem 4.7, given both possible demographically-secret representations one can not only recover an individual's group, but in fact their very identity.

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. 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 [9], that allows preserving utility and controlling unfairness?

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.

REFERENCES

[1] Julia Angwin, Ariana Tobin, and Madeleine Varner. [n. d.]. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. ([n. d.]). https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin Library Catalog: www.propublica.org.

[2] Daniel Berend and Aryeh Kontorovich. 2013. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters* 83, 4 (April 2013), 1254–1259. https://doi.org/10.1016/j.spl.2013.01.023

[3] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *arXiv:1707.00075 [cs]* (July 2017). http://arxiv.org/abs/1707.00075 arXiv: 1707.00075.

[4] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. *arXiv:1810.08810 [cs, stat]* (Oct. 2018). http://arxiv.org/abs/1810.08810 arXiv: 1810.08810.

[5] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [cs]* (Aug. 2018). http://arxiv.org/abs/1808.00023 arXiv: 1808.00023.

[6] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data* 5, 2 (June 2017), 120–134. https://doi.org/10.1089/big.2016.0048 Publisher: Mary Ann Liebert, Inc., publishers.

[7] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz. [n. d.]. Discrimination in Online Advertising A Multidisciplinary Inquiry. ([n. d.]), 15.

[8] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (April 2015), 92–112. https://doi.org/10.1515/popets-2015-0007 Publisher: Sciendo Section: Proceedings on Privacy Enhancing Technologies.

[9] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2017. Guilt-free data reuse. *Commun. ACM* 60, 4 (March 2017), 86–93. https://doi.org/10.1145/3051088

[10] Harrison Edwards and Amos Storkey. 2016. Censoring Representations with an Adversary. *arXiv:1511.05897 [cs, stat]* (March 2016). http://arxiv.org/abs/1511.05897 arXiv: 1511.05897.

[11] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Association for Computing Machinery, Sydney, NSW, Australia, 259–268. https://doi.org/10.1145/2783258.2783311

[12] Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, and Sravya Yandamuri. 2020. Advertising for Demographically Fair Outcomes. *arXiv preprint arXiv:2006.03983* (2020).

[13] James E. Johndrow and Kristian Lum. 2019. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *Annals of Applied Statistics* 13, 1 (March 2019), 189–220. https://doi.org/10.1214/18-AOAS1201 Publisher: Institute of Mathematical Statistics.

[14] Anja Lambrecht and Catherine E. Tucker. 2016. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN Electronic Journal* (2016). https://doi.org/10.2139/ssrn.2852260

[15] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2017. The Variational Fair Autoencoder. *arXiv:1511.00830 [cs, stat]* (Aug. 2017). http://arxiv.org/abs/1511.00830 arXiv: 1511.00830.

[16] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 3384–3393. http://proceedings.mlr.press/v80/madras18a.html

[17] Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. 2017. Provably Fair Representations. *arXiv:1710.04394 [cs]* (Oct. 2017). http://arxiv.org/abs/1710.04394 arXiv: 1710.04394.

[18] Aditya Krishna Menon and Robert C Williamson. [n. d.]. The Cost of Fairness in Binary Classification. ([n. d.]), 12.

[19] Ariana Tobin and Jeremy B. Merrill. [n. d.]. Facebook Is Letting Job Advertisers Target Only Men. ([n. d.]). https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men Library Catalog: www.propublica.org.

[20] Aad W. van der Vaart and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-2545-2

[21] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Sanjoy Dasgupta and David McAllester (Eds.), Vol. 28. PMLR, Atlanta, Georgia, USA, 325–333. http://proceedings.mlr.press/v28/zemel13.html

SUPPLEMENTARY MATERIAL

Proof of Constant Costs Theorem

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 7.1. (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

$$|z| = 2. \tag{44}$$

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 7.2. (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

$$d_H(p, q) \geq k. \tag{45}$$

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

$$\sum_{s \in S} s_i = \sum_{t \in T} t_i. \tag{46}$$

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 7.3. *($(2^{k-1}, 2^{k-2}, 2^k)$-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 4.8.

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,

$$d_H(w, x) \leq d_H(w, z) + d_H(z, x) \tag{47}$$

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,

$$\begin{aligned}\sum_{\{u,v\} \in Z} d_H(u, v) = &\sum_{\{u,v\} \in A} d_H(u, v) \\ &+ \sum_{\{u,v\} \in B} d_H(u, v) \\ &+ \sum_{\{u,v\} \in C} d_H(u, v) \\ &\geq |A|2^{k-2} + |B|2^{k-3} \end{aligned} \tag{48}$$

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,

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

□

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 4.9.

Proof of Lemma 5.5

Proof. We first show that

$$\mathbb{E}[|X_0 - X_1|] = \Omega\left(\sqrt{\text{Var}[X_0]}\right). \tag{49}$$

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

$$\begin{aligned}\mathbb{E}[|X_0 - X_1| \,|\xi] &= \mathbb{E}[|X_0 - \mathbb{E}[X_0] + \mathbb{E}[X_1] - X_1| \,|\xi] &(50)\\ &= \mathbb{E}[X_0 - \mathbb{E}[X_0] \,|\xi] + \mathbb{E}[\mathbb{E}[X_1] - X_1|\xi]\,. &(51)\end{aligned}$$

Observe that

$$\mathbb{E}[X_0 - \mathbb{E}[X_0] \,|\xi] = \mathbb{E}[X_0 - \mathbb{E}[X_0] \,|X_0 > \mathbb{E}[X_0]]\,, \tag{52}$$

and

$$\mathbb{E}[\mathbb{E}[X_1] - X_1|\xi] = \mathbb{E}[\mathbb{E}[X_1] - X_1|\mathbb{E}[X_1] \geq X_1]\,. \tag{53}$$

Since X_0 and X_1 are independent and identically distributed we have

$$\mathbb{E}[\mathbb{E}[X_1] - X_1|\mathbb{E}[X_1] \geq X_1] = \mathbb{E}[\mathbb{E}[X_0] - X_0|\mathbb{E}[X_0] \geq X_0]\,. \tag{54}$$

Therefore

$$\mathbb{E}[|X_0 - X_1|] \geq \mathbb{E}[|X_0 - X_1| \,|\xi] \geq c\mathbb{E}[|X_0 - \mathbb{E}[X_0]|]\,, \tag{55}$$

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 Supplementary Materials) we obtain

$$\mathbb{E}[|X_0 - \mathbb{E}[X_0]|] \geq \sqrt{\frac{\text{Var}[X_0]}{2}}. \tag{56}$$

And therefore

$$\mathbb{E}[|X_0 - \mathbb{E}[X_0]|] = \Omega\left(\sqrt{\text{Var}[X_0]}\right). \tag{57}$$

We now show that

$$\mathbb{E}[|X_0 - X_1|] = O\left(\sqrt{\text{Var}[X_0]}\right), \tag{58}$$

which follows readily since

$$\mathbb{E}[|X_0 - X_1|] \leq \sqrt{\text{Var}[X_0 - X_1]} = \sqrt{2\text{Var}[X_0]}. \tag{59}$$

This completes the proof. □

Probabilistic Inequalities

Theorem 7.4. *(Bretagnolle-Huber-Carol Inequality[20]) Let X be a k-dimensional multinomial random vector with parameters n and $p \in \mathbb{R}^k$, then*

$$\Pr\left[\sum_{i=1}^{k} |X_i - np_i| \geq 2\sqrt{n}\lambda\right] \leq 2^k \exp(-2\lambda^2), \tag{60}$$

$\lambda > 0$.

Theorem 7.5. *(Berend-Kontorovich Inequality[2]) Let X be a binomial random variable with parameters n and p such that $1/n \leq p \leq 1 - (1/n)$, then*

$$\sqrt{\frac{Var[X]}{2}} \leq \mathbb{E}[|X - \mathbb{E}[X]|] \leq \sqrt{Var[X]}. \tag{61}$$