MobilityMirror: Bias-Adjusted Transportation Datasets

Luke Rodriguez[1], Babak Salimi[2], Haoyue Ping[3], Julia Stoyanovich[3], and Bill Howe[1]

[1] Information School, University of Washington, Seattle WA
{rodriglr,billhowe}@uw.edu
[2] Computer Science and Engineering, University of Washington, Seattle WA
bsalimi@cs.washington.edu
[3] College of Computing & Informatics, Drexel University, Philadelphia PA
{hp354,js3735}@drexel.edu

Abstract. We describe customized synthetic datasets for publishing mobility data. Private companies are providing new transportation modalities, and their data is of high value for integrative transportation research, policy enforcement, and public accountability. However, these companies are disincentivized from sharing data not only to protect the privacy of individuals (drivers and/or passengers), but also to protect their own competitive advantage. Moreover, demographic biases arising from how the services are delivered may be amplified if released data is used in other contexts.

We describe a model and algorithm for releasing origin-destination histograms that removes selected biases in the data using causality-based methods. We compute the origin-destination histogram of the original dataset then adjust the counts to remove undesirable causal relationships that can lead to discrimination or violate contractual obligations with data owners. We evaluate the utility of the algorithm on real data from a dockless bike share program in Seattle and taxi data in New York, and show that these adjusted transportation datasets can retain utility while removing bias in the underlying data.

1 Introduction

Urban transportation continues to incorporate new modalities including rideshare [20], bike shares [32], prediction apps for public transportation [12], routing apps for non-motorized traffic [5], and many more. These new services require sharing data between companies, universities, and city agencies to enforce permits, enable integrative models of demand and ridership, and ensure transparency. But releasing data publicly via open data portals is untenable in many situations: corporate data is encumbered with contractual obligations to protect competitive advantage, datasets may exhibit biases that can reinforce discrimination [14] or damage the accuracy of models trained using them [22], and all transportation data is inherently sensitive with respect to privacy [8].

To enable data sharing in these sensitive situations, we advocate releasing "algorithmically adjusted" datasets that *destroy causal relationships between certain sensitive variables* while *preserving relationships in all other cases*.

For example, early deployments of transportation services may favor wealthy neighborhoods, inadvertently discriminating along racial lines due to the historical influence of segregation [1]. Releasing the data "as is" would complicate efforts to develop fair and accurate models of rider demand. For example, card swipe data for public transportation use in Seattle is biased toward employees of tech companies and other large organizations; many neighborhoods predominantly use cash. This bias correlates with race and income, potentially reinforcing social inequities.

Our focus in this paper is to model how these effects manifest in the context of transportation and how to correct for them. We will consider three applications: ride hailing services (using synthetic data), taxi services (using public open data), and dockless bike share services (using sensitive closed data).

We focus on dockless bikeshare services as a primary motivation and a running example. These services allow users to rent a bike through an app with no requirement to return the bike to a fixed station; the bike can simply be left wherever one's trip ends. The City of Seattle began a pilot program for dockless bikes in the Summer of 2017, issuing permits for three different companies to compete in the area (Company A, B, and C). To ensure compliance with the permits, these three companies are required to share data through a third-party university service to enable integrative transportation applications while protecting privacy and ensuring equity. As part of this project, the service produces synthetic datasets intended to balance the competing interests of utility, privacy, and equity. Figure 1 shows a map of the ridership for the pilot program in Seattle and is indicative of the kind of data products generated for transparency and accountability reasons.

There are several potential *sensitive causal dependencies* in these datasets:

- Company A may be moving their bikes into particular neighborhoods to encourage commutes; this strategy could be easily copied at the cost of competitive advantage.
- Company B may be marketing to male riders through magazine ads, leading to a male bias in ridership that could be misinterpreted as demand; data publishers may wish to eliminate the role of gender in the data.
- Company C may be negotiating with the city for subsidies for rides in underserved neighborhoods; they may be disallowed from publicly disclosing information about these subsidies, and therefore wish to remove the relationship between company and demographics.
- Ride hailing and taxi services allow passengers to rate and tip the drivers; gender or racial patterns in tips or ratings may encourage discrimination by drivers and should be eliminated before attempting to develop economic models of tip revenue.

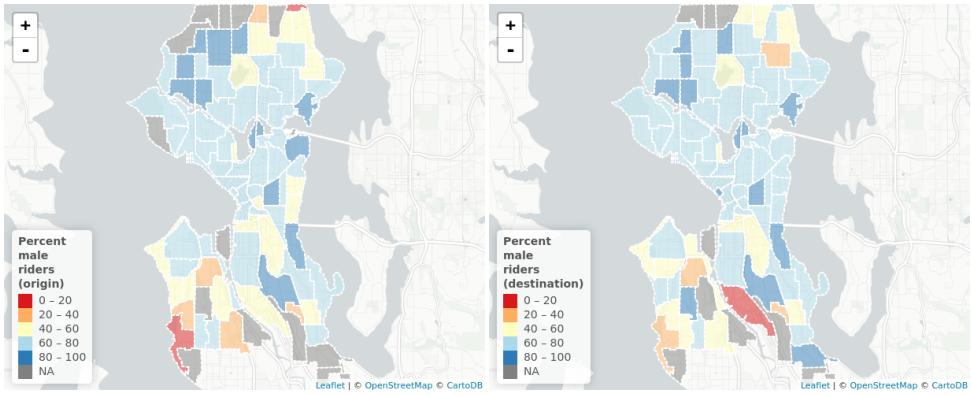

Fig. 1: Percentage of bikeshare trips in Seattle with male riders by origin and destination neighborhoods

In this paper, we develop an approach for adjusting transportation datasets to remove the effects of these sensitive causal relationships while preserving utility for classification and other analysis tasks.

Overview of our approach. Transportation data is frequently released as an *Origin-Destination* (OD) dataset: a set of location pairs representing city blocks, neighborhoods, or other spatial aggregation unit along with the traffic flow between the pair of locations. We augment OD datasets with metadata, such that each tuple represents a histogram bucket corresponding to a unique combination of attributes. For example, the bike share data includes an attribute *gender* with domain $(male, female, other)$ and an attribute *company* with domain (A, B, C) in addition to *origin* and *destination* attributes, each with a domain of 90 neighborhoods around Seattle. A released dataset then might include the tuple $(female, A, Downtown, Ballard, 245)$ indicating that there were 245 trips taken by female riders on bikes owned by company B from Downtown to Ballard during the time period covered by the dataset. These generalized OD datasets are sufficient for a variety of analytics tasks, including modeling demand, evaluating equity, estimating revenue, analyzing ridership trends, and estimating the effect on parking and motorized traffic.

Although these datasets are aggregated, they can still expose sensitive information. Individual privacy is an important concern in data sharing, but we do not focus on it here. In this work, we are primarily interested in other forms of sensitive information encoded in the joint distribution across demographics, neighborhoods, companies, and other attributes. To remove these sensitive patterns, the data publisher specifies a causal relationship between two attributes, written $X \rightarrow Y$, that they wish to eliminate in the adjusted dataset, conditioned on another set of attributes Z. Then the causal repair problem is to set the mutual information between X and Y to zero, conditioned on Z. The conditional attributes Z are important to express realistically complex situations; without

these attributes, degenerate solutions such as scrambling or removing the X or Y attribute altogether would be sufficient, destroying any signal whatsoever.

In our transportation context, our approach corresponds to computing a new distribution of trips over the buckets, one that preserves certain conditional joint probabilities while making other joint probabilities independent. We apply this approach to two real-world datasets of interest: the NYC taxi trip record dataset [25] and dockless bikeshare data from the city of Seattle. The NYC taxi dataset is already available; we choose to evaluate on this dataset to ensure reproducibility. The bikeshare data is legally encumbered and cannot be shared publicly, but we are developing these methods to enable the release of bias-adjusted synthetic datasets to improve transparency.

To evaluate the efficacy of our approach, we show that the distance between the original data and the adjusted data, as measured by multiple appropriate distance metrics, is no greater than would be expected due to sampling variance.

We make the following contributions:

- We describe the bias repair problem for transportation data, which arose from collaborations with companies and city agencies interested in sharing sensitive transportation data.
- We describe a solution for removing a causal dependency (as defined by conditional mutual information) between two attributes in the context of transportation data.
- We evaluate this method on a synthetic rideshare dataset, a real taxi dataset, and a real bikeshare dataset, and demonstrate its effectiveness.
- We discuss generalizations of this approach to other domains, as we well as new potential algorithms to handle specific cases.

We present this method as a research paper, although we are actively seeking legal permission to release a bias-corrected dataset as a companion to the paper.

The rest of this paper is organized as follows: in Section 2 we describe related work in data sharing, causal analysis, and transportation. In Section 3 we present the model of the problem and our proposed algorithm. In Section 4 we describe taxi and bike sharing applications, and explain how our methods are instantiated over the corresponding datasets. In Section 5 we evaluate the algorithm on real and synthetic data. We conclude and discuss possible extensions in Section 6.

2 Related Work

Background on Fairness, Accountability and Transparency (FAT). Recent reports on data-driven decision making underscore that fairness and equitable treatment of individuals and groups is difficult to achieve [2,3,23], and that transparency and accountability of algorithmic processes are indispensable but rarely enacted [4,7,29]. Our approach combines theoretical work relating causality to fairness [15] with practical tools for pre-processing data.

Recent research considers fairness, accountability and transparency properties of specific algorithms and their outputs [9,10,30,33]. However, these approaches are limited by the assumption that no information and no intervention methods are available for the upstream process that generated the input data [16]. Our focus is on developing a practical methodology that improves fairness for these upstream processes, specifically biased transportation data.

Our work can be seen as an extension to methods for generating privacy-preserving synthetic datasets [19,26,31], where the sensitive information that needs to be obscured is about specific relationships between entities rather than simply the presence or absence of the entities themselves.

A common class of approaches to interrogate fairness and quantify discrimination is to use an associative (rather than a causal) relationship between a protected attribute and an outcome. One issue with these approaches is that they do not give intuitive results when the protected attribute exhibits spurious correlations with the outcome via a set of covariates. For instance, in 1973, UC Berkeley was sued for discrimination against females in graduate school admissions,when it was found that 34.6% of women were admitted in 1973 as opposed to 44.3% of men. However, it turned out that women tended to apply to departments with lower overall acceptance rates; the admission rates for men and women when conditioned on department was approximately equal [28]. The data could therefore not be considered evidence for gender-based discrimination.

The importance of causality in reasoning about discrimination is recognized in recent work [6,13,15,24,18,11]. In our work, we focus on discrimination through total and direct effect of a sensitive attribute on an outcome. A comprehensive treatment of discrimination through causality requires reasoning about *path-specific causality* [24], which is difficult to measure in practice, and is the subject of our future work.

3 Model and Algorithm

In this section, we model the bias repair problem, provide some background on causality, and present our solution. We interpret the problem of removing bias from a dataset as eliminating a *causal dependency* between a *treatment attribute* X and an *outcome attribute* Y, assuming *sufficient covariates* $\mathbf{Z}$.

X and Y are conditionally independent given $\mathbf{Z}$ in R, written $(X \perp\!\!\!\perp Y|\mathbf{Z})$, if

$$P_R(X, Y, \mathbf{Z}) = P_R(X, \mathbf{Z})P_R(Y|\mathbf{Z})$$

The strength of a causal link between X and Y is measured by the conditional mutual information between X and Y given $\mathbf{Z}$ [28]. It holds that $(X \perp\!\!\!\perp Y|\mathbf{Z})$ iff the conditional mutual information between X and Y given $\mathbf{Z}$ is zero, written $I(X; Y|\mathbf{Z})$. To remove bias is to enforce $(X \perp\!\!\!\perp Y|\mathbf{Z})$ or, equivalently, to set the conditional mutual information between the treatment and the outcome given the sufficient covariates to zero.

Following an example from the introduction, we can consider the effect of bike share company on gender: one company may market more aggressively

to women, or their bikes may be more difficult for men to ride. This causal dependency warrants removal in various situations. For instance, the company may not want to reveal their marketing strategy, they may not want to be seen as propagating a gender bias, or a model trained on these results may be less generalizable to other cities if this bias is propagated.

Problem Statement: Bias Repair *Given a relation R with a causal dependency ($X \not\perp\!\!\!\perp Y|\mathbf{Z}$), and given a dissimilarity measure Δ between two probability distributions, the* bias elimination problem *is to find R' such that ($X \perp\!\!\!\perp Y|\mathbf{Z}$) while minimizing $\Delta(R, R')$.*

The dissimilarity measure Δ is interpreted as between $P_R(\mathbf{A})$ and $P_{R'}(\mathbf{A})$ (e.g., KL-divergence). We will consider two different distance metrics in Section 5.1: Position-weighted Kendall's Tau (which is rank-sensitive) and Hellinger distance (which is not). We defer a theoretical study of this optimization problem to our ongoing and future work, though we point out a connection to the problem of low-rank matrix approximation [21]. In this paper, we propose an algorithm that directly enforces the independence condition, then show experimentally that the effect on distance is small.

3.1 Background on Causality

We now briefly review causal inference, which forms the basis of our repair algorithm. The goal of causal inference is to estimate the effect of a *treatment* attribute X on an *outcome* attribute Y while accounting for the effects of covariate attributes $\mathbf{Z}$. We compute a *potential outcome* $Y(x)$[27], which represents the outcome if, possibly contrary to fact, the value of X were set to value x. The *causal effect of X on Y* is the expected value of the difference in the potential outcomes $E[Y(x_1) - Y(x_0)]$, called the *average treatment effect(ATE)*. Given a relation with attributes $\mathbf{A}$, ATE can be computed if a sufficient subset of attributes $\mathbf{Z} \subseteq \mathbf{A}$ is available that satisfies the following conditions:

$$\begin{aligned} &\forall \mathbf{z} \in Dom(\mathbf{Z}), \\ &\quad Y(x_0), Y(x_1) \perp\!\!\!\perp X \mid \mathbf{Z} = \mathbf{z} && \text{(Unconfoundedness)} \\ &\quad 0 < \Pr(X = x_1 \mid \mathbf{Z} = \mathbf{z}) < 1 && \text{(Overlap)} \end{aligned}$$

The first condition requires that the outcome is independent of the treatment given any values for the other attributes. The second condition ensures that the probability does not degenerate to 0 or 1 for any values of the other attributes.

If these conditions are met, the following holds:

$$\text{ATE} = \sum_{\mathbf{z} \in Dom(\mathbf{Z})} \left(\mathbb{E}[Y|X = x_1, \mathbf{Z} = \mathbf{z}] - \mathbb{E}[Y|X = x_0, \mathbf{Z} = \mathbf{z}]\right) \Pr(\mathbf{Z} = \mathbf{z}) \quad (1)$$

where $Dom(\mathbf{Z})$ is the domain of the attributes $\mathbf{Z}$.

It can be shown that the ATE of X on Y is zero iff $I(X; Y|\mathbf{Z}) = 0$. Therefore, we are justified in using the conditional mutual information $I(X; Y|\mathbf{Z})$ to quantify the strength of a causal link between X and Y given $\mathbf{Z}$.

ATE quantifies the total effect of X on Y, which can be separated into direct effects and indirect effects (those that are mediated through other attributes). In this paper, we ignore this distinction for expository purposes, and leave generalizing the method to account for this distinction to future work.

3.2 Algorithm

We propose a simple algorithm to compute an approximate solution to our problem. The algorithm is based on the intuition that $(X \perp\!\!\!\perp Y|\mathbf{Z})$ holds in R' iff the joint probability distribution $\Pr_{R'}(\mathbf{A})$ admits the following factorization, based on the chain rule:

$$P_{R'}(\mathbf{A}) = P_{R'}(X\mathbf{Z})P_{R'}(Y|\mathbf{Z})P_{R'}(\mathbf{U}|XY\mathbf{Z}) \tag{2}$$

where $\mathbf{U} = A - XY\mathbf{Z}$.

This factorization will form the basis of our algorithm, but there is a complication: We want to restrict R' to include only the active domain of R rather than the full domain. The reason is that transportation datasets are typically sparse; there are many combinations of attributes that do not correspond to any traffic (e.g., bike rides from the far North of the city to the far South). We assume R is a bag; it may contain duplicates. For example, there may be multiple trips with the same origin, destination, and demographic information. Under this semantics, we express our algorithm in terms of *contingency tables*.

A contingency table over a set of attributes $\mathbf{X} \subseteq \mathbf{A}$, written $\mathcal{C}_R^{\mathbf{X}}$, is simply the count of the number of tuples for each unique value of $\mathbf{x} \in Dom(\mathbf{X})$. That is, $\mathcal{C}_R^{\mathbf{X}}$ corresponds to the result of the query `select` $\mathbf{X}$`, count(*) from` R `group by` $\mathbf{X}$. More formally, a contingency table over $\mathbf{X} \subseteq \mathbf{A}$ is a function $Dom(\mathbf{X}) \to \mathbb{N}$

$$\mathcal{C}_R^{\mathbf{X}}(\mathbf{x}) = \sum_{t \in R} \mathbb{1}[t[\mathbf{X}] = \mathbf{x}]$$

$t[\mathbf{X}]$ represents the tuple t projected to the attributes $\mathbf{X}$, and $\mathbb{1}$ is the indicator function for the condition $t[\mathbf{X}] = \mathbf{x}$. The contingency table over all attributes in R is an alternative representation for the bag R itself: Given $\mathcal{C}_R^{\mathbf{A}}$, we can recover R by iterating over $Dom(\mathbf{A})$. In practice, this step is not necessary, as $\mathcal{C}$ is implemented as a k-dimensional array.

Using contingency tables, we can compute a new joint probability distribution over $\mathbf{A}$ as

$$P_R(\mathbf{A} = a) = \frac{\mathcal{C}_R^{\mathbf{A}}(a)}{|R|}$$

Algorithm 1 uses these ideas to construct the desired relation R' from the marginal frequencies of R, enforcing Equation 2 by construction. It can be shown that the KL-divergence between $P_R(\mathbf{A})$ and $P_{R'}(\mathbf{A})$ is bounded by $I(X;Y|\mathbf{Z})$. That is, the divergence of R' from R depends on the strength of the causal dependency between X and Y. If the causal dependency is weak, Algorithm 1 will have no significant effect on the dataset. We will evaluate the effects experimentally in Section 5.

Algorithm 1: Enforcing Conditional Independence

Input: An instance R with $\mathbf{A} = XY\mathbf{ZU}$ in which $(X \not\perp\!\!\!\perp Y|\mathbf{Z})$
Output: An instance R' in which $(X \perp\!\!\!\perp Y|\mathbf{Z})$

for $xy\mathbf{zu} \in R$ **do**
  $numerator \leftarrow \mathcal{C}_R^{X\mathbf{Z}}(x\mathbf{z})\mathcal{C}_R^{Y\mathbf{Z}}(y\mathbf{z})\mathcal{C}_R^{XY\mathbf{ZU}}(xy\mathbf{zu})$
  $denominator \leftarrow |R|\mathcal{C}_R^{\mathbf{Z}}(\mathbf{z})\mathcal{C}_R^{XY\mathbf{Z}}(xy\mathbf{z})$
  $\mathcal{C}_{R'}^{\mathbf{A}}(xy\mathbf{zu}) \leftarrow \mathbf{Round}(\frac{numerator}{denominator})$
return R' associated with $\mathcal{C}_{R'}^{\mathbf{A}}$

Table 1: Processed NYC taxi data

orig_lon	orig_lat	dest_lon	dest_lat	pickup_time	distance	tip	driver_freq	count
-74.0	40.7	-74.0	40.7	night	medium	high	low	6074
-74.0	40.7	-74.0	40.7	night	medium	low	low	2844
-73.9	40.7	-73.9	40.7	day	low	high	medium	16
-73.9	40.7	-73.9	40.7	morning	low	high	low	14
-73.9	40.7	-74.0	40.7	morning	low	high	high	3

4 Applications and Datasets

In this section we describe two real datasets to which we apply our methodology and an overview of how both datasets were processed for use in our evaluation.

NYC Taxi Data The NYC taxi trip record dataset released by the Taxi & Limousine Commission (TLC) [25] contains trips for 13,260 taxi drivers during January 2013, with pick-up and drop-off location as (lat,lon) coordinates and other information including trip distance and tip amount. We used this particular release of the data because medallion numbers were no longer made available after this release. We first removed transportation records with missing values, such as records with unknown pick-up or drop-off locations or missing tip amount. We then categorized trip distance into low, medium, and high, with about 1/3 of the trips falling into each category. Tip amount was categorized into low and high, with high tip corresponding to at least 20% of the fare amount. Note that the original dataset has tip amount information only for rides that were paid by a credit card, and so we only consider these trips in the paper. Lastly, drivers were categorized into low, medium, and high frequency drivers. Table 1 shows an example of the data after aggregation, with the count of each instance represented in the `count` column.

Dockless Bikeshare The bike data includes rides from $197,049$ distinct riders between June 2017 and May 2018 across three different companies. Each rider is identified via a unique rider id for each company, and the start and end location of each trip is projected to one of 94 neighborhoods in the Seattle area. Trip

information is joined with rider information from survey responses, indicating their gender and whether or not they use a helmet.

Data Processing and Aggregation We pre-processed both datasets to make them compatible with our approach. First, the time in both is precise up to the second. Since our model assumes categorical attributes, we map time to four buckets: morning (5am - 9am), day (9am - 3pm), evening (3pm - 7pm), and night (7pm - 5am). Additionally, each individual driver/rider was classified into one of three categories by the number of trips they made, as recorded in the dataset. The top 1/3, who made the most trips, are designated `heavy`, the bottom 1/3 are designated `light`, and the rest are designated `medium`.

5 Experiments

In this section, we first outline our evaluation metrics, and then present experiments to consider whether the error introduced by our bias-repair method is comparable to the error introduced by natural variation. Recall that we wish to remove the causal dependency between X and Y. If there is no correlation between these attributes, then the repair process will not change the weights significantly. If there is a strong correlation, then the process will force the mutual information to zero while preserving the distribution of the other attributes.

We consider three situations: synthetic data simulating extreme situations (Section 5.2), real datasets representing bike and taxi data (Section 5.3), and the same real bike and taxi data, but aggregated post hoc to simple origin-destination pairs (Section 5.4). The experiments in each of these situations can be summarized by the choice of treatment (X), outcome (Y) and covariate (Z) attributes, $X \rightarrow Y|Z$, as follows:

1. Synthetic: $gender \rightarrow rating|\{origin, destination\}$
2. Bike: $company \rightarrow gender|\{start_nhood, end_nhood, time_of_day, helmet\}$
3. Taxi: $distance \rightarrow tip|\{orig_lon, orig_lat, dest_lon, dest_lat\}$

5.1 Evaluation Metrics

Our goal is to remove the effect of the given relationship without destroying the utility of the resulting dataset. The proposed method would not be viable if it altered the distribution of traffic "too much." To define "too much," we a) compute the distance between the original dataset and the adjusted dataset, and b) compare this distance with the distances associated with a set of bootstrap samples of the original dataset. If the distance with the adjusted dataset falls within the distribution of the bootstrap samples, we conclude that the adjustment is small enough to still produce a useful dataset.

To compute distances, we consider two different metrics: one that is rank-sensitive, and one that is not. To measure rank-sensitive distance, we sort the buckets by trip count in descending order before and after the repair. We then

use position-weighted Kendall's tau [17] to compare the two resulting rankings. Kendall's tau counts the number of pair-wise position swaps between a ground truth ranking and an experimental ranking. Position-weighted Kendall's tau incorporates a weighting function, usually to assign more importance to swaps that happen closer to the beginning of the ranked list.[4] This measure is appropriate in our domain, because a) transportation analysts and engineers are primarily interested in the conditions associated with the heaviest traffic flows, and b) transportation datasets are inherently sparse.

The weighting function we consider is harmonic: Given position i in a ranking, the weight is $\frac{1}{i}$. We also considered an exponential weighting function, since traffic patterns tend to follow an exponential distribution, but that weighting function was potentially too generous to our method: The first few positions were all that mattered.

To measure distance independently of rank and position, we use Hellinger distance. This measure is an f-divergence closely related to the Bhattacharyya distance that obeys the triangle inequality, and is defined as follows: Let p, q be two probability distributions over the same set of attributes $\mathbf{X}$, and define the Bhattacharyya Coefficient $BC(p, q)$ to be $\sum_{x \in \mathbf{X}} \sqrt{p(x)q(x)}$. Then the Hellinger distance is $H(p, q) = \sqrt{1 - BC(p, q)}$.

Table 2: Results of evaluation metrics across all experiments

Dataset	PWKT 2.5%	Mean	97.5%	**Exp. Result**	Hellinger 2.5%	Mean	97.5%	**Exp. Result**
Synth. - uncorrelated	1.47	2.93	4.39	**0.159**	0.075	0.076	0.076	**0.00079**
Synth. - correlated	1.35	2.44	3.54	**3.18**	0.072	0.073	0.074	**0.42**
Bike - all categories	1.49	2.34	3.18	**1.53**	0.084	0.085	0.086	**0.15**
Taxi - all categories	1.37	2.88	4.39	**1.21**	0.14	0.15	0.15	**0.042**
Bike - aggregated	0.84	1.39	1.93	**1.37**	0.029	0.030	0.030	**0.024**
Taxi - aggregated	0.27	0.81	1.36	**0.40**	0.044	0.047	0.051	**0.0020**

Table 2 presents results for both position-weighted Kendall's tau (PWKT) and Hellinger distance in each of our experiments. The experimental result for Algorithm 1 is in bold, and the other columns summarize the distribution of the bootstrap samples. Figure 2 visualizes these results. Each experiment is represented by three bars. The light bar on the left shows the distribution of distances from the bootstrap procedure: the top of the bar represents the 97.5 percentile, the next line represents the mean, and lowest line represents the 2.5 percentile. We visualize the distribution as a bar to emphasize that the measure is a distance, such that a lower bar is always better. The dark bar in the center

[4] Many methods for comparing ranked lists have been proposed. We opt for a measure in which identity of the items being ranked (histogram buckets) is deemed important. This is in contrast to typical IR measures such as NDCG or MAP, where item identity is disregarded, and only item quality or relevance scores are retained.

is the experimental result. The final bar on the right represents a baseline test of assigning every trip a random X value as a strategy of enforcing $I(X;Y) = 0$.

Overall, we can see that the error introduced by our algorithm is usually significantly less than the error one can expect from sampling, suggesting that the method is viable for correcting bias while retaining utility.

The expected variation is clearly visible for the case of PWKT, but for the Hellinger distance it is small compared to the magnitude of the metric, and is nearly impossible to distinguish precisely. The *Correlated* and *Bike* columns for the Hellinger distance stand out as significant outliers.

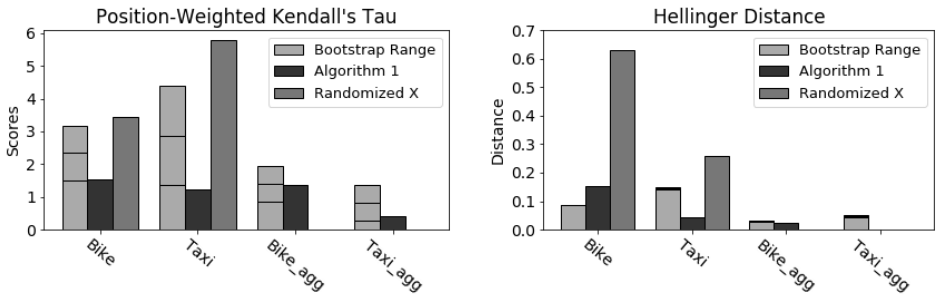

Fig. 2: Expected bootstrap variation (left), experiment outcome (center), and baseline comparison (right) for each of the four experiments on real data. Lines in the bar denote the 2.5%, mean, and 97.5% values for the distribution.

5.2 Synthetic Ride Hailing Data

For the synthetic experiments, the task is to remove the causal influence of gender on rating, simulating the situation where a data publisher does not want to unintentionally encourage discrimination [14].

To generate the synthetic data, we use neighborhood-level trip data from the dockless bikeshare to simulate a realistic distribution of traffic among neighborhoods. Then, we assign each individual trip a gender at random from $\{m, f, o\}$ representing male, female, or other. The *no correlation* experiment assigns ratings according to a pre-defined distribution independent of gender, while the *gender correlated* experiment uses three different distributions, one for each gender value, to simulate a strong correlation. In both of these experiments, our simulated repair is to remove the effect of gender on rating, conditional on the origin and destination neighborhoods.

We expect that the uncorrelated case should have minimal effect on the data, since there is no causal dependency to eliminate. For the strongly correlated case, we expect the error to be significant.

Comparing the synthetic data experiments in Table 2, we see that there is a change in the order of magnitude of the effect when the repair is acting on a relationship with a strong underlying correlation. When applied to synthetic data with no correlation structure, we find that values of both position-weighted Kendall's tau and Hellinger distance fall well below the range of error introduced by bootstrap sampling. However, in the correlated case when there was in fact a strong relationship, position-weighted Kendall's tau jumped to the upper extreme of the bootstrap range, and Hellinger distance far exceeded this range. This result indicates that the repair is causing a more drastic change in the gender correlated case than in the case with no correlation, as we would expect.

Position-weighted Kendall's tau still falls within the bootstrap variation for the correlated case, which can be explained by the fact that certain neighborhood origin-destination pairs carry a disproportionate amount of the traffic in the dataset, so this relationship is preserved. The full magnitude of the change is better observed through the Hellinger distance in Table 2, which grows an order of magnitude beyond the bootstrap variance in the gender correlated experiment.

5.3 Real-World Bike and Taxi Datasets

In the bike experiment we remove the influence of company on gender using the dockless bikeshare data described in Section 4. In this experiment, we are considering the situation where companies are releasing data to support traffic research, but do not want to expose any latent gender bias that may be attributable more to marketing efforts than to sexism. The relationship between company and gender is conditional on origin, destination, and whether or not the rider uses a helmet. In other words, *only the effect of company should be removed, not the overall pattern of gender on ridership.*

In the taxi experiment we investigate the effect of a repair on the taxi data from Section 4, in which we remove the influence of distance on tip amount, conditional on origin and destination. The situation we consider is a behavioral economic analysis of tipping patterns, but we want to completely remove the influence of distance. Simply normalizing by distance is not enough, as the joint distribution between, say, time of day, distance, and tip amount can be complex. Moreover, certain neighborhood origins and destinations may generate higher tips or lower tips in ways that interact with distance traveled. For example, long east-west trips at certain times may be relatively short, but generate higher tips.

In both cases, we see that the calculated position-weighted Kendall's tau and Hellinger distance in Table 2 fall close to the expected variation from bootstrap samples, with the exception of the Hellinger distance for the bike share data, which is about twice this baseline. This anomaly helped us discover a data ingest error upstream from our algorithm: gender information was only properly included for one company, while the other two had two different default values. As a result, there was an unrealistically high correlation between company and gender. The order of values was still largely preserved by Algorithm 1, as seen in Figure 2, since there are significantly more trips from one company than from the others, but the structural change results in a high Hellinger distance.

Taken along with the taxi data, this reaffirms that Algorithm 1 behaves as expected: it induces larger changes when there is a high degree of correlation in the relationship chosen for treatment.

5.4 Aggregated Origin-Destination Data

In our experiments so far we considered all possible fine-grained buckets in the dataset. For example, the trip count associated with {`UDistrict, Downtown, Female, Helmet, Morning`} appears as a bucket. We also consider a coarser aggregated view of this data, grouping buckets by origin and destination and aggregating over gender, helmet, and time. The motivation is that, in many situations, only origin-destination counts are important, and also that our method may unfairly benefit on a fine-grained dataset: if we preserve the distribution of the top few origin-destination pairs, we will also preserve the distribution of a large number of finer-grained buckets that divide these origin-destination pairs by gender, helmet and time. We run the same experiments and metrics as before, but this time grouping by origin and destination.

When aggregating as described, we see in Figure 2 that the baseline (right column) for each of these experiments has a value of 0. This is because origin and destination were not included in X or Y, and any repair that only takes into account the relationship between X and Y does not impact the other direct relationships in the dataset. For the results of Algorithm 1 (center), the Hellinger distance falls below the expected variation for both datasets, while the position-weighted Kendall's tau falls in the bottom half of the expected range of variation. We therefore conclude that Algorithm 1 preserves both order and structure of real aggregated data at least as well as a bootstrapped sample, given these particular correlation structures.

6 Conclusions and Future Work

Data sharing is emerging as a critical bottleneck in urban and social computing. While risks associated with privacy have been well-studied, data owners and data publishers must also be selective about the patterns they reveal in shared data. Biases in the underlying data can be reinforced and amplified when used to train models, leading to not only poor quality results but also potentially illegal discrimination against protected groups, causing a breach of trust between government and companies.

In this paper, we have considered the bias-correction problem — an important pre-processing step in releasing data that is orthogonal to privacy.

We interpret the need to repair unintended or unrepresentative relationships between variables prior to data release as related to causal inference: the conditional mutual information between two variables is a measure of the strength of the relationship. We propose an algorithm that interprets the frequencies of trip events as a probability distribution, then manipulates this distribution to eliminate the unwanted causal relationship while preserving the other relationships.

We show that this procedure produces expected behavior for synthetic datasets representing extreme cases, and has only a modest impact in real datasets: the distance between the original data and the adjusted data falls within the bounds of natural variation of the original data itself.

Going forward, we aim to generalize this approach to other domains, distinguish between direct and indirect causal effects, and explore new algorithms that can better balance the tradeoff between utility and causal relationships. Our broader vision is to develop a new kind of open data system that can spur data science research by generating safe and useful synthetic datasets on demand for specific scenarios, using real data as input.

References

1. Amazon doesn't consider the race of its customers. should it? *Bloomberg*, 2016.
2. J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: Risk assessments in criminal sentencing. *ProPublica*, May 23, 2016.
3. S. Barocas and A. Selbst. Big data's disparate impact. *California Law Review*, 2016.
4. R. Brauneis and E. P. Goodman. Algorithmic transparency for the smart city. *Yale Journal of Law & Technology*, forthcoming.
5. A. M. Brock, J. E. Froehlich, J. Guerreiro, B. Tannert, A. Caspi, J. Schöning, and S. Landau. Sig: Making maps accessible and putting accessibility in maps. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, page SIG03. ACM, 2018.
6. A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE SP*, pages 598–617, 2016.
7. A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *PoPETs*, 2015(1):92–112, 2015.
8. Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 2013.
9. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012.
10. 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, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268, 2015.
11. 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*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.
12. B. Ferris, K. Watkins, and A. Borning. Onebusaway: results from providing real-time arrival information for public transit. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1807–1816. ACM, 2010.
13. S. Galhotra, Y. Brun, and A. Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2017, Paderborn, Germany, September 4-8, 2017*, pages 498–510, 2017.

14. Y. Ge, C. R. Knittel, D. MacKenzie, and S. Zoepf. Racial and gender discrimination in transportation network companies. Working Paper 22776, National Bureau of Economic Research, October 2016.
15. N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.
16. K. Kirkpatrick. It's not the algorithm, it's the data. *Commun. ACM*, 60(2):21–23, Jan. 2017.
17. R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 571–580, 2010.
18. M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.
19. W. Lu, G. Miklau, and V. Gupta. Generating private synthetic databases for untrusted system evaluation. In *IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014*, pages 652–663, 2014.
20. S. Ma, Y. Zheng, and O. Wolfson. Real-time city-scale taxi ridesharing. 27:1782–1795, 07 2015.
21. I. Markovsky. *Low rank approximation: algorithms, implementation, applications*. Springer Science & Business Media, 2011.
22. D. A. McFarland and H. R. McFarland. Big data and the danger of being precisely inaccurate. *Big Data & Society*, 2(2):2053951715602495, 2015.
23. MetroLab Network. First, do no harm: Ethical guidelines for applying predictive tools within human services. `http://www.alleghenycountyanalytics.us/`, 2018. [forthcoming].
24. R. Nabi and I. Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.
25. NYC Taxi and Limousine Commission. TLC trip record data, 2018. `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`,accessed on June 2, 2018.
26. H. Ping, J. Stoyanovich, and B. Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 42:1–42:5, 2017.
27. D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
28. B. Salimi, J. Gehrke, and D. Suciu. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1021–1035. ACM, 2018.
29. L. Sweeney. Discrimination in online ad delivery. *Commun. ACM*, 56(5):44–54, 2013.
30. R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, pages 325–333, 2013.
31. J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. PrivBayes: private data release via Bayesian networks. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 1423–1434, 2014.

32. Y. Zhang, T. Thomas, M. Brussel, and M. van Maarseveen. Expanding bicycle-sharing systems: lessons learnt from an analysis of usage. *PLoS one*, 11(12):e0168604, 2016.
33. I. Zliobaite. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.*, 31(4):1060–1089, 2017.