Tackling Peer-to-Peer Discrimination in the Sharing Economy

Naman Goel
naman.goel@epfl.ch
Swiss Federal Institute of Technology
Lausanne, Switzerland

Maxime Rutagarama
maxime.rutag@alumni.epfl.ch
Swiss Federal Institute of Technology
Lausanne, Switzerland

Boi Faltings
boi.faltings@epfl.ch
Swiss Federal Institute of Technology
Lausanne, Switzerland

ABSTRACT

Sharing economy platforms such as Airbnb and Uber face a major challenge in the form of *peer-to-peer discrimination* based on sensitive personal attributes such as race and gender. As shown by a recent study under controlled settings, reputation systems can eliminate social biases on these platforms by building trust between the users. However, for this to work in practice, the reputation systems must themselves be non-discriminatory. In fact, a biased reputation system will further reinforce the bias and create a vicious feedback loop. Given that the reputation scores are generally aggregates of ratings provided by human users to one another, it is not surprising that the scores often inherit the human bias. In this paper, we address the problem of making reputation systems on sharing economy platforms more fair and unbiased. We show that a game-theoretical incentive mechanism can be used to encourage users to go against common bias and provide a *truthful* rating about others, obtained through a more careful and deeper evaluation. In situations where an incentive mechanism can't be implemented, we show that a simple post-processing approach can also be used to correct bias in the reputation scores, while minimizing the loss in the useful information provided by the scores. We evaluate the proposed solution on synthetic and real datasets from Airbnb.

CCS CONCEPTS

• **Information systems** → *Trust*; *Incentive schemes*; *Reputation systems*; • **Human-centered computing** → *Reputation systems.*

KEYWORDS

Reputation Systems, Sharing Economy, Fairness, Incentive Design

ACM Reference Format:
Naman Goel, Maxime Rutagarama, and Boi Faltings. 2020. Tackling Peer-to-Peer Discrimination in the Sharing Economy. In *12th ACM Conference on Web Science (WebSci '20), July 6–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3394231.3397926

1 INTRODUCTION AND RELATED WORK

Since the creation of eBay in 1995, the basic idea of peer-to-peer sharing of goods and services has led to many successful commercial platforms on the web. This way of distributing goods and services is often called as the sharing economy. The field saw a dazzling boom in the early 2010's when the popularity of Uber and Airbnb began to soar. Today several sharing economy platforms are operational on the web in diverse areas such as travel, real-state, transport, labor, finance and technology etc. The platforms offer an attractive alternative to both the 'producers' and the 'consumers' over the traditional ways of doing business due to being more easily accessible, sustainable and decentralized in nature. However, several recent studies have highlighted some very important ethical challenges faced by these platforms. For example, Edelman et al. [4] found that Airbnb booking requests from the researchers (posing as guests) were 16 percent less likely to be accepted when the researchers made the requests from guests accounts with distinctively African American names relative to the case when they used identical guests accounts with distinctively white names. Similarly, Edelman and Luca [5] found that prices of properties on Airbnb offered by black hosts tend to be significantly lower than their white counterparts, even while keeping other relevant factors constant. We conjecture that less demand or trust for properties offered by black hosts is one of the reasons for the lower prices. Beyond Airbnb, Ge et al. [8] have observed racial discrimination by Uber drivers via more frequent cancellations against passengers when the researchers used African American sounding names for passenger accounts. Thus, the discrimination exists both ways. Hosts and drivers (providers) racially discriminate among guests and passengers (consumers) and vice-versa. It is deeply concerning that the existing social biases are finding their way into web based platforms too. A combination of biased human feedback and large scale algorithmic decision-making on the web can cause further social segregation of historically disadvantaged groups. Thus, the problem needs an urgent attention and solution.

Abrahao et al. [1] (from Airbnb and Stanford) conducted an extensive user study on real Airbnb users and claimed that reputation systems offset the real world social biases by building trust between different users. The design of this user study was motivated from the concept of trust based investment games in economics. This is indeed a very positive finding. Trust between the users is the fundamental reason why Airbnb works in the first place [9]. However, it may be noted that even though the study was conducted with real Airbnb users, the reputation scores used in the study were generated synthetically. The reputation systems must themselves be non-discriminatory for them to actually work in the expected manner. A reputation system that discriminates against people based on race or gender will only further reinforce the bias. Unfortunately, the reputation systems on these platforms are often discriminatory towards different races and genders. This was analyzed in great detail by Hannák et al. [11] for freelancer marketplaces like taskrabbit.com and fiverr.com. We found a similar trend on Airbnb too in our study. The findings are not very surprising

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WebSci '20, July 6–10, 2020, Southampton, United Kingdom
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7989-2/20/07…$15.00
https://doi.org/10.1145/3394231.3397926

because reputation systems are based on aggregating the rating provided by users (humans) to one another and human society has a long history of racial bias and discrimination. Unfortunately, it is a non-trivial problem to make reputation system non-discriminatory because the goal of the reputation systems is really to discriminate between users. However, this discrimination must be based on relevant attributes and not on sensitive attributes like race or gender. Thus, the objective is to make reputation systems racially non-discriminatory while retaining the other useful information they provide. This problem is similar to the problem of making machine learning systems non-discriminatory, where there is a trade-off between increasing the discriminative power (classification accuracy) of the classifiers and reducing racial discrimination in their decisions at the same time. The latter problem has received a lot of attention recently [2, 3, 7, 10, 12, 14, 18–20]. However, the solutions are specific to machine learning (mostly classification algorithms) and don't apply to reputation systems where humans are directly responsible for the reputations scores.

Contributions. In this paper, we propose two solutions to make reputations systems on sharing economy platforms more fair and non-discriminatory. The first solution is to incentivize users (example guests on Airbnb) to find potentially high quality service providers (example hosts on Airbnb) from the disadvantaged group, evaluate them through deeper inspection (by using their service) and provide a truthful review of the service. We show that a game theoretic peer-consistency mechanism called the Peer Truth Serum for Crowdsourcing (Radanovic et al. [17]) can use the knowledge about the sensitive attribute of the service providers to ensure desired incentive compatibility in this scenario. This solution differs from the idea of offering explicit incentives just to explore the unexplored services (for example, see Hirnschall et al. [13]). We provide the incentives for exploration while truthfully rating the service, ensuring that the incentive mechanism doesn't cause a "reverse discrimination" and doesn't make it easier for the disadvantaged group to get business. If the design of the platform doesn't allow an incentive mechanism to be implemented, we propose a second solution. This solution applies to any reputation system irrespective of the reputation aggregation algorithm and the rating behavior of the users. We transform the aggregated reputation scores, such that the transformed scores are non-discriminatory to desired level while ensuring as little loss in their informativeness as possible. We model the problem as a constrained convex optimization problem and learn optimal transformation parameters that minimize information loss while respecting the constraints on the covariance between transformed scores and sensitive attribute(s).

2 AIRBNB CASE STUDY

2.1 Dataset

As discussed in the introduction, the phenomenon of peer-to-peer discrimination on the sharing economy platforms is already well-documented in the literature. Since we would need a real dataset to evaluate our proposed solutions and the datasets used in prior studies are not publicly available, we collected a new dataset for this paper. We used the data available on Inside Airbnb [1], which is "an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world". We collected data for the listings in the New York City. The attributes that are of interest to us in this data include the aggregate rating (reputation scores) of the hosts, the prices of the listings, the profile pictures of the hosts and several other characteristics of the listings (for example number of bedrooms, bathrooms, guests, accommodates, min nights, reviews, reviews per month, location coordinates etc). We also add two labels to each listing: the ethnicity of the host and whether the listing is in a black majority neighborhood. To get the ethnicity of the host, we used the profile picture of the hosts and a face recognition library API called Kairos[2]. Kairos API, given the URL of an image, returns information about the people detected in the image, including a confidence score between 0 and 1 for five possible ethnicities: asian, black, white, hispanic or other. For each listing in our dataset, we take the ethnicity as the one having the maximal confidence score from Kairos results, and we remove the listing if this confidence score is not higher than a threshold (fixed at 0.7). This thresholding ensures that we filter out the listings for which Kairos couldn't detect the ethnicity of the hosts with enough confidence. To determine whether a given listing is in a black majority neighborhood or not, we used the coordinates of the listing and an additional piece of data (also available on Inside Airbnb) that contains the coordinates of the boundaries of the neighborhoods (for example, Harlem, Queens Village, Jamaica etc) in New York City. Using this information and a spatial analysis library in Python (Shapely), we were able to determine the exact neighborhood of each of the listings. We then used census data to determine whether a given neighborhood is black majority or not. This classification is also available online.[3]. After all these pre-processing steps, we finally get a dataset of 8218 listings on Airbnb from New York City. Based on host ethnicity, 5716 listings are from white hosts and 2502 from non-white hosts (due to comparatively small proportions of other ethnicities in the dataset, we merged all non-white ethnicities). 3748 listings are in black-majority neighborhoods and 4470 are in other neighborhoods. Note that the imbalance in the dataset is a feature of the real-world. We had also collected similar datasets from Amsterdam, Geneva and San Francisco but the datasets were even more imbalanced (very few listings from minority groups) and hence, we skip discussion of those datasets in this paper. It may also be noted that Kairos also allows us to find the age and gender of the hosts. We skip discussion about the distribution of these attributes as they don't lead to any interesting findings w.r.t. discrimination.

2.2 Data Analysis

As discussed in the introduction, there have also been user studies on Airbnb and Uber that go beyond just correlation analysis and present more compelling evidence of discrimination. But in this paper, we will restrict ourselves to correlation analysis. In our data, there are two main attributes about a listing (and the corresponding host) that we focus on: the average rating (reputation scores) of the hosts and the prices of the listings. The average ratings are direct controlled by guests while prices are indirectly controlled by guests

[1] https://insideairbnb.com

[2] https://kairos.com

[3] https://en.wikipedia.org/wiki/List_of_African-American_neighborhoods

(due to lower demand and trust). Thus, if we find a significant difference in the values of these attributes for different ethnic groups, then it can be a potential case of bias. Table 1 shows the difference in average prices of the listings and the average rating of white and non-white hosts. Table 2 shows the difference in the listings of the hosts in black majority and other neighborhoods.

Table 1: Ratings and prices for different host ethnicity

	White Hosts	Others	Relative Difference
Avg. Price	$139.12	$105.51	31%
Avg. Rating	93.97	92.62	1.5%*

Table 2: Ratings and prices for different neighborhoods

	Others	Black Majority	Relative Difference
Avg. Price	$155.66	$98.64	57%
Avg. Rating	94.02	93.20	0.9%*

*These values are actually more significant than they seem. Ratings on Airbnb are almost always bigger than 85 (0.1 quantile is at 84.7), so the range of the ratings as shown above is not really 0-100 and the relative difference could be much higher after scaling (around 18% and 10% respectively if we consider the range to be between 85 and 100).

To further confirm the bias, we perform a regression analysis on the prices of the listings and ratings of the hosts (as target variables) with all the observable features that we could get from Inside Airbnb, including ethnicity of the host and neighborhood type. A similar approach was followed in [11] for confirming bias on taskrabbit and fiverr. The regression results obtained using Python's statsmodels library (linear model, OLS) are shown in Table 3 and 4. The 'ceof' columns shows the linear relationship between observed variables and the price (or ratings) and the '$P > |t|$' column shows the p-values for the relationship. This confirms that even after accounting for the observable features, ethnicity of the host (with positive correlation for white hosts) and the majority ethnicity of the neighborhoods (with negative correlation for black majority areas) have a statistically significant effect on the ratings and prices. Another point to note in Table 3 is that, as expected, ratings also have a statistically significant positive effect on prices. While the platform itself has no direct control over the prices, it can definitely design better reputation systems which would then affect prices as well. In the rest of the paper, we will focus only on the ratings.

Remark. We will be using the example of guests discriminating among hosts throughout the paper, but the solutions proposed in the paper are more general and apply for tackling discrimination in the reverse direction as well (most platforms like Airbnb and Uber have two-way reputation systems but it is more difficult to collect data about the reputation scores of guests and passengers).

3 BIAS FREE RATING ELICITATION

Online reputation systems involve two main steps. In the first step, the users provide ratings and in the second step, the platform aggregates the ratings into reputation scores. In this section, we intervene in the first step and discuss how an incentive mechanism can elicit fair ratings from users.

Table 3: Regression Analysis for Price

	coef	std err	t	P>|t|
const	-69.8484	19.962	-3.499	0.000
accommodates	26.7803	1.646	16.268	0.000
bathrooms	31.6218	4.180	7.565	0.000
bedrooms	16.0144	3.346	4.786	0.000
beds	-7.0650	2.845	-2.483	0.013
guests	8.5204	1.756	4.852	0.000
min nights	0.4624	0.102	4.547	0.000
reviews	-0.0011	0.054	-0.020	0.984
reviews/month	-7.0454	1.204	-5.852	0.000
rating	0.9372	0.204	4.588	0.000
black majority area	-55.7336	3.326	-16.755	0.000
white host	14.3567	3.608	3.979	0.000

Table 4: Regression Analysis for Ratings

	coef	std err	t	P>|t|
const	93.4629	0.313	298.216	0.000
accommodates	-0.1241	0.090	-1.375	0.169
bathrooms	-0.6018	0.226	-2.659	0.008
bedrooms	0.1657	0.181	0.916	0.360
beds	-0.2936	0.154	-1.912	0.056
price	0.0027	0.001	4.588	0.000
guests	0.2543	0.095	2.680	0.007
min nights	-0.0152	0.005	-2.772	0.006
reviews	-0.0100	0.003	-3.432	0.001
reviews/month	0.3918	0.065	6.030	0.000
black majority area	-0.5371	0.182	-2.943	0.003
white host	1.0296	0.195	5.291	0.000

Incentive Mechanism Design Goals.

(1) Users should be encouraged to try the services of high quality individuals belonging to the disadvantaged class.
(2) Users should provide truthful ratings about the received quality of service. Providing a truthful rating about the service is important because we don't want to offer unconditional benefits for any individual belonging to any class.

While there may be several ways to achieve the first goal alone, achieving the second goal together with the first goal is a hard problem because it also requires that user must reveal their private information (the quality of service that they received from the individual) truthfully, even though there is no way to verify what the user is saying is indeed true. The latter problem is known as the problem of "information elicitation without verification" and there is a rich literature on mechanisms for this problem. The mechanisms are commonly referred to as the "peer-consistency" mechanisms [6]. The examples of these mechanisms include the original peer-prediction method of Miller et al. [15] and the Bayesian Truth Serum of Prelec [16]). The broad idea in these mechanisms is to "match" the information provided by different users and reward the users based on agreement between the two pieces of information. We next show that a state-of-the-art peer-consistency mechanism called the Peer Truth Serum for Crowdsourcing (Radanovic

et al. [17]) achieves both design goals listed above, if it has access to the sensitive attribute of the individuals on the platform.

3.1 Preliminaries and Notation

Let the individuals providing services on a platform be categorized into two classes, based on the value z of their sensitive attribute Z. For example, on Airbnb, hosts can be categorized into black hosts and white hosts. In the case of freelancer platforms like taskrabbit, fiverr (Hannák et al. [11]), workers can be categorized into male and female workers. The sensitive attribute is not necessarily required to be binary one and there can also be multiple sensitive attributes. The discussion in the paper can also be extended to these general cases. The users of the platform, who take the services of these individuals, have private prior beliefs about the underlying quality of the services provided by individuals belonging to different classes. The prior belief of a user may be different for different classes. For example, a user may believe that male workers are able to provide better service or that white hosts provide better accommodation. We formalize beliefs of users as probability distributions about the quality of service. Let the quality of service be expressed using a discrete signal Q that can take values in $\{1, 2, \ldots, k\}$, 1 being worst and k being best. Then, prior belief of a user about the quality of service provided by an individual belonging to class $Z = z$ is given by $P_z(Q = q)$. Once the user personally observes the service quality q' offered by an individual Ψ (for example after hiring Ψ or staying in the accommodation offered by Ψ), he updates his belief about this particular individual Ψ. This is expressed by his posterior belief $P_\Psi(Q = q|q')$: given that he received a service of quality q' himself, the probability that any other user on the platform will receive a service of quality q from the same individual Ψ.

3.2 The Peer Truth Serum for Crowdsourcing (Radanovic et al.)

Let r denote the rating given by a user i to an individual Ψ belonging to class $Z = z$ and let r' denote the rating given by another (randomly selected) user j to the same individual. Then, the Peer Truth Serum rewards user i with τ_i given by:

$$\tau_i = \beta(z)\left[\frac{\mathbb{1}_{r=r'}}{R_{iz}(r)} - 1\right]$$

Here $\mathbb{1}_{r=r'}$ is an indicator function which returns 1 if the ratings r and r' match and 0 otherwise. $R_{iz}(r)$ is the relative frequency of r in the ratings received by all individuals of class $Z = z$, excluding the ratings given by user i. More formally, $R_{iz}(r) = \frac{num_{iz}(r)}{\sum_{\bar{r}\in\{1,2,\ldots,k\}} num_{iz}(\bar{r})}$, where $num_{iz}(r)$ is a function that counts occurrences of r in the ratings of all individuals (except i) of class $Z = z$. $\beta(z)$ is a strictly positive scaling constant for the class $Z = z$ such that the constant for the disadvantaged class is sufficiently bigger than the constant for the other class. Rewards can also be calculated by taking average by matching the ratings with multiple other users instead of single user (to reduce variance).

3.3 Belief Update Assumption

We will make a weak and standard assumption (the self-predicting assumption [17]) about the way users update their belief after they observe a quality of service. We assume:

$$\frac{P_\Psi(Q = q|q)}{P_z(Q = q)} > \frac{P_\Psi(Q = q'|q)}{P_z(Q = q')} \qquad \forall\, q, q' \in \{1, 2, \ldots K\}$$

The assumption says that the relative change in posterior (over the prior) about what quality other users will observe from individual Ψ is the highest for the quality that the user himself observed. The assumption is easiest to understand in binary (good or bad quality) settings. If the user herself observed a good service, his belief about others receiving a good service from this individual doesn't decrease (or remain exactly same) as compared to his prior.

3.4 Game-Theoretic Properties

(1) **Truthful Equilibrium[17].** Under the self-predicting assumption on workers' beliefs (which can be heterogeneous and unknown), the mechanism induces a truthful Bayes-Nash equilibrium: if other users submit truthful ratings, it is the best strategy for any user to submit truthful rating. Further, the expected reward in the truthful equilibrium is strictly positive.

Proof Sketch: A rational user seeks to maximize the expected reward. The numerator in the first term of expected reward is the posterior belief of the user about another user receiving a certain quality of service given his own observation and the denominator converges to his prior belief about that quality of service offered by a random individual from that class. Then, truthful equilibrium follows from self-predicting assumption. A formal proof can be found in [17]. It holds even if the rewards are scaled by constants.

(2) **Robustness to Collusion[17].** The mechanism ensures that truthful equilibrium is not just an equilibrium but the most profitable equilibrium. So collusion strategies are not profitable. For example, a simple strategy in which users may always submit the same rating (irrespective of true quality) so that their rating always match gives zero expected reward.

Proof Sketch: If everyone gives same rating irrespective of what quality they actually observed, then while numerator is always 1, the denominator (relative frequency of that rating) is also always 1 (net reward is 0). More generally, for ratings provided randomly (independent of the true quality), numerator and denominator converge to same quantity [17].

(3) **Higher Reward for Truthful Ratings in the *Disadvantaged Class* but No Free Ride.** The mechanism gives higher reward if the users (try the service and) truthfully rate good quality individuals from disadvantaged class but doesn't incentivize giving good rating for a bad service.

Proof Sketch. The scaling factor β is higher for the disadvantaged class but due to the truthfulness property of the mechanism, the higher scaling factor only helps when the provided rating is also truthful.

Remarks. Assuming that the rewards given by the mechanism are small compared to the cost paid by the users and the disutility of actually receiving a bad service, the mechanism only incentivizes

the users to actively search for individuals within the disadvantaged class that are very likely to offer high quality service (for example, based on photographs /presentation of the service). It is true that in this way, the mechanism also benefits individuals from the disadvantaged class who already have high ratings, but if this is not desired, one can re-define the class of individuals taking into account other attributes also (for example prior number of reviews) and apply the mechanism at a finer level of class definition. For example, on Airbnb example, 'super hosts' status already makes this distinction. Finally, we note that the scaling constant β can be dynamically changed and made equal for the two groups once it has served its purpose (i.e. when there is no disparity on the platform). The mechanism then continues to incentivize truthful reporting, leading to a sustainable long-term fairness on the platform.

3.5 Sensitive Attribute Information

Our mechanism uses the information about the sensitive attribute (class) of the individuals to calculate $R_{iz}(r)$ using the ratings of all individuals across a class $Z = z$ and also to scale the rewards with class dependent scaling constants $\beta(z)$. This information is required to ensure that the mechanism works in the desired way.

To give a concrete example, imagine that the $R_{iz}(r)$ for some r denoting a good rating is 0.8 for the disadvantaged class and is 0.9 for the other class. Remember that $R_{iz}(r)$ is estimate of the prior belief. Let's assume that after a user receives a good service from an individual, her posterior belief that another user will also receive a good service from the same individual, increases by 0.02. Thus, it becomes 0.82 if the individual is from the disadvantaged class and 0.92 if the individual is from the other class. In this example, the user gets an expected reward of $\beta(z)\left[\frac{0.82}{0.8}-1\right] = 0.025\beta(z)$ or $\beta(z)\left[\frac{0.92}{0.9}-1\right] = 0.022\beta(z)$ depending on the class of the individual being rated. Let's further assume for simplicity that the same happens in case of a bad service also i.e. her posterior belief that another user will also receive a bad service from the same individual given that she herself received bad service, increases by 0.02. Thus, it becomes 0.22 if the individual is from the disadvantaged class and 0.12 if the individual is from the other class. In this example, the user gets an expected reward of $0.1\beta(z)$ or $0.2\beta(z)$ depending on the class.

Now imagine that we instead use an $R_i(r)$ calculated from the ratings received for all individuals irrespective of their class, and a class independent scaling constant β. $R_i(r)$ may be the average of 0.9 and 0.8, for example. Using this $R_i(r)$ violates all three properties enumerated in section 3.4. The first two properties are violated because a common $R_i(r)$ is no longer an estimate of the different prior beliefs for the two classes and the self-predicting assumption can not be used to guarantee truthfulness. The third property is violated because the scaling constant is no longer different for the difference classes. In fact, not using the sensitive attribute information in the mechanism may cause even more discrimination. The rewards no longer encourage exploring good service providing individuals from the disadvantaged class but on the contrary, may discourage doing so. In the above example, it is easy to see that truthfully rating a good service gets a negative reward of $\beta\left[\frac{0.82}{0.85}-1\right]$ for the disadvantaged class and a positive reward of $\beta\left[\frac{0.92}{0.85}-1\right]$ for the other class. On the other hand, truthfully rating a bad service gets a positive reward of $\beta\left[\frac{0.22}{0.15}-1\right]$ for the disadvantaged class and a negative reward of $\beta\left[\frac{0.12}{0.15}-1\right]$ for the other class. Indeed, it is not necessary that a similar difference in rewards for the two classes will always be observed (depending on different priors and belief update parameters) but the example clearly shows that it is possible that the bias may be reinforced if the sensitive attribute information is not used in the mechanism. Hence, it is important to use this information to achieve the desired outcome using the incentive mechanism. This is similar to using sensitive attribute in fair machine learning (at training and/or prediction time).

4 BIAS CORRECTION

While the incentive mechanism proposed in the previous section is an attractive method to make reputation systems fair, it is possible that the design and business model of some sharing economy platforms may not permit the implementation of such a mechanism since many platforms may be reluctant to pay for reviews. In such cases, an ideal solution would be to somehow estimate the bias of the users and discount their opinions according to their bias parameters while aggregating the ratings into the reputation scores. However, such a solution can only work if:

- There is enough data available about every user. This means that every user must provide enough ratings across classes.
- The data from users follows a consistent probabilistic distribution so that their biased behavior can be learned. In our case, this requires that users provide ratings in a consistent way.

Unfortunately, most users on the web provide very few ratings making it hard to make any inference about their bias parameters. Further, it is difficult to model human rating behavior using a simple probabilistic model. Hence, we take an alternative approach and propose a "post-aggregation" correction technique.

4.1 Post-Aggregation Transformation

We apply a transformation on the aggregated reputation scores of all individuals such that the transformed scores are non-discriminatory, while ensuring that the information loss due to transformation is minimum. More formally, let $x = \{x_1, x_2, \ldots x_n\}$ be the originally aggregated reputation scores of individuals $\{1, 2, ...n\}$ and let $\tilde{x} = \{\tilde{x}_1, \tilde{x}_2, \ldots \tilde{x}_n\}$ be their transformed scores. Let $l(\tilde{x}, x)$ be a loss function measuring the amount of information lost due to the transformation and let $d(\tilde{x}, z)$ be a function measuring the discrimination in the transformed scores, $z = \{z_1, z_2, \ldots z_n\}$ being the values of the sensitive attribute. Then, we have the following problem:

$$\begin{aligned} \text{minimize} \quad & l(\tilde{x}, x) \\ \text{subject to} \quad & d(\tilde{x}, z) \leq \delta \end{aligned} \tag{1}$$

Here, $\delta \geq 0$ is the allowed threshold of discrimination in the transformed scores. l and d can be chosen according to the domain of application and the computational considerations. In our paper, we make the following assumptions: (1) l is the mean squared error(MSE), which is a common choice for measuring difference in real valued ratings (for example, to evaluate recommendation systems algorithms), (2) d is the absolute value of covariance between the transformed scores and the sensitive attribute values, and (3)

$\tilde{x}_i = a \cdot x_i + b$. With these assumptions, we get a convex optimization problem in a and b, which can be solved efficiently using existing tools. The sensitive attribute doesn't have to be binary and it is easy to accommodate non-binary categorical sensitive attributes using one-hot representation of the sensitive attribute (a common trick used in data analysis and machine learning).

Remarks. The choice of covariance as a proxy to measure bias was explored by Zafar et al. [18] in context of machine learning classifiers. In our settings, correlation would be a more appropriate metric than covariance due to scale invariance, but it makes the problem non-convex. Nevertheless, as we will show, even covariance turns out to give good performance in our experiments. Finally, we note that instead of using identical transformation parameters a, b for all individuals, one could define $\tilde{x}$ to have individual specific parameters a_i, b_i (at the cost of increasing the number of optimization parameters). It is also possible to consider more complex transformation functions (for example, $\tilde{x}_i = a \cdot x_i^2 + b \cdot x_i + c$). However, it may not always be useful if $d(\tilde{x}, z)$ is still measured using covariance (covariance captures only the linear dependence between two variables). For example, even with $\tilde{x}_i = a \cdot x_i + b$ and covariance as the measure of bias, it is easy to show mathematically that the closed form solution for a is 1 in our optimization problem. The same is true for coefficients of higher order terms. This was also observed in our experiments. Thus, we only discuss simple transformations of the form $a \cdot x_i + b$ in this paper. It remains an interesting future work to explore whether other transformation functions $\tilde{x}$ *(together with more advanced measures $d(\tilde{x}, z)$ of discrimination)* can achieve better performance, while keeping the problem convex.

4.2 Range Scaling

On many platforms (including Airbnb), aggregated reputation scores always lie within a fixed range $[L, U]$ (for example, [0,5] or [0,100]). It is fairly easy to address this in our proposed solution. One natural fix is to include additional constraints on the parameters a, b in the optimization problem such that $L < \tilde{x}_i = a \cdot x_i + b < U$ for any $L < x_i < U$ and constants L, U. An even simpler approach is to apply a range scaling on the transformed scores (after optimization) so that the transformed scores lie in desired range $[l, u]$.

$$\tilde{x}_i^{scaled} = (u - l)\frac{\tilde{x}_i - \min(\tilde{x})}{\max(\tilde{x}) - \min(\tilde{x})} + l$$

In our experiments, we set $l = \min(x), u = U$. This ensure that the the minimum value of the transformed scores doesn't go below the minimum value of the original reputation scores. It may be noted that the scaling is a constant linear function applied to all scores; hence, the covariance between the scaled transformed scores and the sensitive attribute will stay the same as it was before scaling, and it doesn't alter the discrimination removal achieved by the constrained optimization step.

Remark. Assuming the conclusions of the user study conducted by [1], a one time correction in reputation scores should bring fairness on the platform by building trust between the users. In a less optimistic (and perhaps more realistic) scenario, the correction can be applied only at infrequent intervals, eventually leading to an ideal setting where no more corrections are required and reputation scores are fair by default.

4.3 Experiments

We implemented the above approach in Python (using Scipy's Sequential Least Squares Programming) and tested it on our Airbnb dataset. The results presented here are for the case when host ethnicity was assumed to be the sensitive attribute but similar trends were observed when neighborhood ethnicity was assumed to be the sensitive attribute. The value of δ was set to 10^{-5}.

Table 5: MSE and Covariance after the transformation

	$MSE(x, \widetilde{x})$	$cov(\widetilde{x}, z)$
Before Transformation	0	0.228
After Transformation	0.246	10^{-5}

Table 5 shows that the covariance between the reputation scores and the ethnicity was reduced to $1e-05$ as specified by the constraint (δ) and MSE increased to 0.246. While it is certainly good that covariance is now close to 0, it is not clear whether the increase in MSE is acceptable or not. Even a naive transformation technique (for example which assigns random reputation scores to individuals independent of their true reputation) could also achieve a zero covariance but that would clearly not be an acceptable solution. Thus, we perform a regression analysis on the transformed reputation scores (exactly as we did in Section 2). Table 6 shows that the p-value corresponding to the host ethnicity is now 0.803 >> 0.05, which means ethnicity is now an insignificant feature, while p-values for other relevant remain unchanged. This shows that our technique retained the desired information while removing discrimination. We note that if there are multiple sensitive attributes, our optimization framework can easily accommodate multiple constraints (one for each sensitive attribute). For example, if both host ethnicity as well as neighborhood's majority ethnicity are specified as sensitive attributes in the constraints, then the scores transformed using the two constraints would show no relation with both attributes.

We now show some additional experimental results on synthetic datasets.

- **Power Law Distribution.** We assumed that the sensitive attribute can now take 3 different values (for example black, white and asian hosts) and generated 5000, 10,000 and 15,000 reputation scores (one for each synthetic host) in the range 0-10 for these three classes. We generated the scores using

Table 6: Regression Analysis for Transformed Ratings

	coef	**std err**	**t**	**P>|t|**
const	94.2129	0.313	300.610	0.000
accommodates	-0.1241	0.090	-1.375	0.169
bathrooms	-0.6018	0.226	-2.659	0.008
bedrooms	0.1657	0.181	0.916	0.360
beds	-0.2936	0.154	-1.912	0.056
price	0.0027	0.001	4.588	0.000
guests	0.2543	0.095	2.680	0.007
min nights	-0.0152	0.005	-2.772	0.006
reviews	-0.0100	0.003	-3.432	0.001
reviews/month	0.3918	0.065	6.030	0.000
black majority area	-0.5371	0.182	-2.943	0.003
white host	-0.0486	0.195	-0.250	0.803

power-law distribution (with parameters 3, 6 and 10 respectively) to closely model the distribution observed in Airbnb data. Table 7 shows the results of the transformation for $\delta = 0.01$.

Table 7: Synthetic Data (Power Law Distribution)

	$MSE(x, \widetilde{x})$	$cov(\widetilde{x}, z)$
Before Transformation	0	0.264
After Transformation	0.117	0.01

- **Normal Distribution.** Similar observations were made when the data was generated using truncated normal distributions ($10,000$ scores from $\mathcal{N}(5,1)$ and $10,000$ from $\mathcal{N}(8,2)$). Instead of presenting duplicate trends, we instead show some other interesting observations on the data generated using normal distribution. Figure 1 shows that the our threshold constant δ in the optimization problem provides the platform direct control over the extent to which reputation scores can be altered, leading to different values of covariance and MSE. The flat parts of the curves show that if δ is set to a value equal to or higher than the covariance of the original scores, then as expected our algorithm does not transform the scores and return the original scores. Figure 2 shows that the proposed technique scales linearly with the number of reputation scores to be transformed. We transformed upto 1M reputation scores (equal number of samples from two truncated normals) in under 4 minutes on a normal PC.

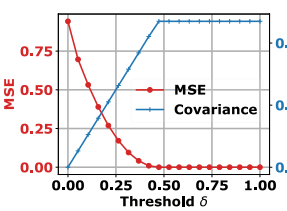

Figure 1: Controlling δ

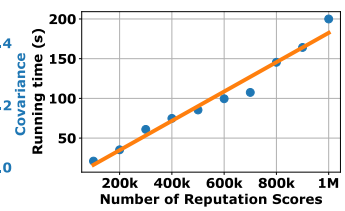

Figure 2: Running Time

5 CONCLUSION

In this paper, we considered the problem of making reputation systems on the sharing economy platforms more fair and proposed two solutions: an incentive mechanism and a bias correction technique. The incentive mechanism encourages users to try the service of individuals belonging to the disadvantaged class and at the same time also elicits truthful ratings about the quality of service received. This ensures that disadvantaged class doesn't receive unconditional benefit and the aggregated reputation scores compensate for any disparate benefit in the long term. Eventually, when individuals have reputation scores that truly reflect their quality irrespective of their class, the incentive mechanism can be easily modified through the scaling constant to stop offering different incentives for different classes, and it can continue to offer incentives for truthful reporting. The bias correction is also meant to be a similar short-term intervention, albeit a more directly controlled one. Even if the platform uses the bias correction solution instead of the payment mechanism, the mechanism can still be used to provide feedback to the raters through artificial currency or points.

There also remain several open questions. Usually, the platforms not only display the aggregated reputation scores but also each of ratings given by the users. On one hand, this further helps in emphasizing that an incentive mechanism is an ideal way to make reputation systems fair but on the other hand, it also means that a post-aggregation bias correction technique loses its utility since the bias is corrected only for the aggregated reputation scores. There is no trivial way to hide biased ratings without access to sufficient data to estimate biased behavior of some of the users. Another open problem is to elicit truthful textual reviews or to filter biased ones.

REFERENCES

[1] Bruno Abrahao, Paolo Parigi, Alok Gupta, and Karen S Cook. 2017. Reputation offsets trust judgments based on social biases among Airbnb users. *Proceedings of the National Academy of Sciences* (2017), 201604234.
[2] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).
[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.
[4] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics* 9, 2 (April 2017), 1–22. http://www.aeaweb.org/articles?id=10.1257/app.20160213
[5] Benjamin G Edelman and Michael Luca. 2014. Digital discrimination: The case of airbnb. com. (2014).
[6] Boi Faltings and Goran Radanovic. 2017. Game Theory for Data Science: Eliciting Truthful Information. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11, 2 (2017), 1–151.
[7] 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*. ACM, 259–268.
[8] Yanbo Ge, Christopher R. Knittel, Don MacKenzie, and Stephen Zoepf. 2016. *Racial and Gender Discrimination in Transportation Network Companies.* Working Paper 22776. National Bureau of Economic Research. https://doi.org/10.3386/w22776
[9] Joe Gebbia. 2016. How Airbnb designs for trust. *TED. com* (2016).
[10] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-Discriminatory Machine Learning through Convex Fairness Criteria. In *Proceedings of AAAI Conference on Artificial Intelligence*.
[11] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1914–1933.
[12] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*.
[13] Christoph Hirnschall, Adish Singla, Sebastian Tschiatschek, and Andreas Krause. 2018. Learning user preferences to incentivize exploration in the sharing economy. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
[14] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *8th Innovations in Theoretical Computer Science Conference (ITCS)* (2017).
[15] Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51, 9 (2005).
[16] Dražen Prelec. 2004. A Bayesian truth serum for subjective data. *Science* 306, 5695 (2004), 462–466.
[17] Goran Radanovic, Boi Faltings, and Radu Jurca. 2016. Incentives for Effort in Crowdsourcing using the Peer Truth Serum. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 4 (2016), 48.
[18] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *26th International World Wide Wed Conference (WWW)*.
[19] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
[20] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. *Neural information processing systems* (2017).