The Price of Fairness in Location Based Advertising

Chris Riederer
Columbia University
New York, NY 10027
mani@cs.columbia.edu

Augustin Chaintreau
Columbia University
New York, NY 10027
augustin@cs.columbia.edu

ABSTRACT

Firms use massive amounts of personal data to decide which advertisements to show to an individual, raising concerns of fairness and algorithmic bias. Previous work has proposed techniques to make machine learning more fair through awareness of the protected attributes of user data. However, these studies have either focused on specific tasks, been primarily theoretical, or have ignored the highly important domain of location-based advertising.

In this work, we present an empirical analysis of the impact of fairness on advertising revenue using a real world example: location based ad personalization for users of Instagram. We empirically analyze the potential for inadvertent discrimination among gender and race in location-based systems, additionally showing the impact of location representation on fairness. Furthermore, we apply fairness techniques to analyze how revenue is affected when both individual and group fairness guarantees must hold. Though this work is a grounding for research into fairness in location-based ads, our methodology applies to more general advertising tasks.

ACM Reference format:
Chris Riederer and Augustin Chaintreau. 2017. The Price of Fairness in Location Based Advertising. In *Proceedings of ACM RecSys conference, Como, Italy, August 2017 (FATREC'17),* 5 pages.
https://doi.org/10.18122/B2MD8C

1 INTRODUCTION

Every day, personal data becomes more broadly available and its use in analytics and advertising clearly generates large sums of wealth. What is perhaps less clear is how tools to prevent discrimination against vulnerable populations can keep up with the growth in algorithmic decision-making based on this personal data. Here we focus on informing what can *practically* be done to guarantee fairness when location data is used in targeted advertising. We choose this application for multiple reasons: It is increasingly common as location-based personalization reaches a large part of the population and it is hard to evade. As we empirically demonstrate, mobility data has great benefits but raises many concerns in the way it is currently used. Perhaps more importantly, we show that many of the hardest challenges previously addressed in theoretical terms can be quantified in this scenario. For instance, this brings us to revisit questions like "What constitutes a practical definition of fairness?", "What should we know or trust about those exploiting the data?", "What is the gain we lose when some definition of fairness must be enforced?"

Let us describe a motivating example where disparate outcomes in targeted advertising is undesirable. For instance, consider a website advertising hiring opportunities to users; its goal is to optimize for relevance as long as disparate outcomes among genders and races are avoided. Why would such a system pose new challenges? First, previously proposed solutions focus on reconciling learning and fairness for *specific tasks for a single party* [1, 2, 10, 11]. For instance, how to increase loan repayment while satisfying equality of treatment or opportunity. In contrast, data providers interact with myriad third parties each leveraging data for different learning tasks. Second, as is commonly the case for online data providers, data about individuals are sparse and naturally represented in high dimensions. This contrasts with solutions designed to learn from a few structured features available for all users, such as exam scores. Additionally, leveraging data at large scale invariably means that computational complexity becomes a severe constraint, so each optimization to reconcile fairness with accuracy will rely on efficient approximation.

These challenges, however, do not imply that no solutions can be found to deploy fair targeting. The direction we examine here is to transform location data before they are used to train and target individuals. If the transformation and targeting satisfies some conditions (see background below), then fairness can be guaranteed for *any* task. As we demonstrate, much of the gains from targeting is preserved. For concreteness and simplicity, we focus in this short article on the simplest transform where details of mobile data are remove by grouping records into larger location cells.

2 BACKGROUND

In our work, we use the definitions of "Fairness Through Awareness" [3], distinguishing between fairness at an individual level and at a group level, which we describe in detail below.

Individual fairness. The main principle is that similar people should see similar outcomes. More rigorously, we consider a classification setting where individuals (denoted by the set V) are mapped to probability distributions over outcomes A. For simplicity, throughout this work we will say each outcome is the decision of whether to show either a generic or targeted ad, and denote these outcomes as $A = \{0, 1\}$ with $A = 1$ corresponding to the decision to show a targeted ad and $A = 0$ a generic ad instead. The space of probability distributions defined on A is $\Delta(A)$. From our point of view, a machine learning algorithm using data from the mobile ad-network defines a mapping $M : V \to \Delta(A)$. A difference score between individuals is denoted by $d : V \times V \to [0; 1]$ and a difference score between probability distributions is D. Throughout this paper, without loss of generality, as a choice to measure the distance between probabilistic outcomes we will use D_{TV}, the distance of total variation (equivalent to one half the $\mathbb{L}_1$ norm) though others can be used. It is defined as: $D_{TV}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$.

FATREC'17, August 2017, Como, Italy

DOI: 10.18122/B2MD8C

Given these definitions, an algorithm is *individually fair* if for all individuals x and y, we have

$$D_{TV}(M(x), M(y)) \leq d(x, y) \tag{1}$$

Intuitively, this says that an advertising system must show similar sets of ads to similar users, and mathematically, this means that that an algorithm mapping users to distributions over outcomes must be Lipschitz continuous.

[3] shows that it is possible in polynomial time to find a mapping M that is both individually fair and maximizes a linear objective function (such as expected revenue) using a linear program.

Group fairness. In contrast to individual fairness, [3] defines two groups of users S and T as having statistical parity up to bias ε when:

$$D_{TV}\ (E_S[M], E_T[M]) \leq \varepsilon \tag{2}$$

where E_S and E_S denotes the expectation of ads seen by an individual chosen uniformly among S and T. This definition implies that the difference in probability between two groups of seeing a particular ad will be bounded by ε. Note that individual fairness does not imply group fairness, and vice versa. A natural question is: "When can both individual fairness and statistical parity be achieved simultaneously"? To guide the design of a mobile platform one can use the following result introducing $d_{EM}(S, T)$, the Earth Mover's Distance [6] between S and T.

Theorem 1. *Given a distance $d \leq 1$, among all algorithms M that are individual fair, for any subsets of users S, T we have*

(i) *It always holds that* $D_{TV}\ (E_S[M], E_T[M]) \leq d_{EM}(S, T)$,

(ii) *there exists M such that* $D_{TV}\ (E_S[M], E_T[M]) = d_{EM}(S, T)$.

3 DATA DESCRIPTION

To understand the important trade-offs facing advertising platforms, we collected a behavioral dataset linked to race and gender information. We obtained publicly available data from Instagram, a popular image sharing social network. Instagram data includes behavioral data such as locations and short texts of describing activities as well as the photos themselves which provide information through the use of computational vision techniques.

3.1 Methodology

We gathered metadata (such as time of photo, URL of image, tags, location, etc.) for all photographs of a "root" user, Kevin Systrom, the founder of Instagram. We then randomly sampled user profiles from those who had commented or liked his photos and gathered their metadata. We repeated this process, randomly sampling user IDs of those commenting or liking photos of any crawled profiles, obtaining the metadata of 115,796,284 for 260,389 different profiles. Systrom is a popular Instagram presence (1.4M followers) and a wide variety of users comment on his photos, seemingly to communicate with the platform, making him a good starting point for a random crawl. No images were downloaded from Instagram.

Location. Of our 115 million photo information dataset, 16,537,404 were geotagged for 162,549 users. In order to study advertising that micro-targets small granularity locations, we narrowed our focus to two major United States cities, New York City and Los Angeles, a typical practice. Using only photos located in the bounding boxes of those two cities, we created two subsets: New York had 22,300 users with 707,265 photos and Los Angeles had 20,724 users with 776,065 photos.

Dataset	Number Users	Number Checkins	Labeled Gender	Labeled Race
New York	22,300	707,265	10,388	902
Los Angeles	20,724	776,065	9,748	851

Table 1: Overview of dataset used in study.

Tags. Like other social networks, Instagram users label their content with "hashtags", which label topics for the photo, make photos more easily searchable, or let the user express him- or herself. As we discuss in a later section, we use these tags later as part of our location-based advertising model.

3.2 Labeling

Labeling gender. To label our the gender of the users in our dataset, we applied the methodology of Mislove et al. [4]. We obtained the number of babies born by name, gender, and year of birth in the United States via Social Security data[1], assigning a gender to users with a first name for which there were both at least 50 births and 95% of recorded births were one gender. Out of our entire dataset of 260 thousand users, this labeled 92,935 profiles (35%). In our New York City subset, 10,388 were labeled with gender, 5,471 female and 4,917 male. In Los Angeles, 9,748 users labeled with gender: 4,965 female and 4,783 male.

Race labeling. We labeled the race of profiles based on face recognition software, similar to prior work [5]. The Face++ API (`www.faceplusplus.com`) recognizes faces in images, additionally providing demographic information, labeling the race of users from one among Asian, Black, and Caucasian. Although we did not download any photos, our metadata included publicly accessible URLs of images, which we could pass to the Face++ API. We ran this software on the first 500 photographs of a subset of our New York and Los Angeles users, labeling a profile with a binary race classification (Caucasian or minority) that appeared most frequently in their photographs. This labeled 902 users in our New York dataset; 746 labeled Caucasian and 156 from minorities, and 851 users in Los Angeles; 710 Caucasian and 141 minority.

Evaluation with manual labeling. To provide ground truth validation of our more scaled labeling techniques, two research assistants labeled a randomly selected subset of 200 profiles for gender and race. After filtering for private, deleted, or business profiles, 194 profiles remained. Of our 194 human-labeled profiles, 86 users had first names recognized by our methodology. Of these, 84 out of 86 (97%) agreed, giving us high confidence in the precision of our gender labeling approach. For race labeling, our computational vision approached agreed with human labelers 89.7% of the time. comparable to other works that report that Face++ has high levels of accuracy for race labeling.

4 MOBILE ADVERTISING MODEL

In order to analyze the trade off between fairness and revenue, we model a location-based advertising system using our dataset. We focus on this domain due to its importance (38% of all smartphone advertising used location targeting in 2016), and its potential for

[1] Available at https://www.ssa.gov/oact/babynames/limits.html

discrimination as location is highly sensitive and often correlates with sensitive traits such as race or income [8]. We simulate a system with the following problem: Given a user's locations from previous check-ins, predict what topics a user will be interested in. Such a prediction could allow a service to better target ads.

4.1 User and Location Representation

We represent individuals in terms of their visits to different locations. We map locations to an index j. Each user is represented as an array, with index j set to 1 if the user has checked in at location j and a 0 otherwise. In our original dataset, locations for each photo are latitude-longitude pairs, and here we discretize these by truncating these coordinates to a certain level of prevision. In different analyses we vary this precision to study how fairness and revenue is impacted by granularity of location representation. Using fewer digits implies a lower granularity, which is better for privacy but less specific and hence likely less useful for advertisers. We vary the cell sizes from 0 decimal places (*e.g.* (-74., 40.) is a cell; cells have sides of length roughly 111km) to 4 places (e.g. (-73.9989, 40.7245) is a cell; cells have sides of roughly 10m). We additionally conducted our analysis representing users with a histogram of frequencies of visits to each location as opposed to binary representations, but the results were similar and we omit them due to space.

4.2 Interest Prediction

After defining how users are represented, we use these feature to predict if a user is interested in several topics, utilizing Instagram's hashtags for ground truth. Hashtags, used on several platforms such as Instagram and Twitter, are ways for users to associate topics with their post. Examples include a user tagging a picture of food with "#food" or of himself with "#selfie". We use three different tags: #fashion, #travel, and #health.

We trained a model predicting a user's likelihood to post each of the three tags using a user's location visits as features and whether or not they had used a tag as labels. To avoid overfitting we regularized each model using ridge regression (i.e. $\mathbb{L}_2$ penalty) and conducting three way cross validation, picking the parameter that maximized peformance on the training set. All training was conducted using the scikit-learn python package.

4.3 Performance and Revenue Estimation

We evaluate our models in two ways: in traditional machine learning terms and for their ability to improve revenue in an advertising simulation. We use AUC as a metric to understand our classifier performance due to its standard acceptance and our class distributions being highly skewed. For all three tags and both cities, AUC is 0.5 at the broadest granularity, meaning our model is no better than random guessing. However, as the number of digits increases, so does AUC. In NYC, our classifiers have AUCs of 0.82, 0.92, and 0.65 for fashion, health, and travel, respectively, and in LA, we report AUCS of 0.83, 0.92, and 0.68.

Moving beyond classifier performance, we estimated the impact of granularity on revenue. Earlier, we distinguished between generic and targeted advertisements. Based on estimates generated from the Facebook ad tool[2], we said that the cost per click (advertiser revenue) for a targeted ad was \$2 and the revenue for a generic ad was \$1. In our model, a generic ad always generates revenue, and a targeted ad only generates revenue if the user is indeed interested in a topic, and so the system will only show a targeted ad to a user if the expected revenue justifies the risk of receiving no revenue. Using this model, a predictor using the finest granularity of 4 digits generated \$1021, \$994, and \$906 in revenue for fashion, travel, and health, respecitvely, over a baseline of displaying generic dislay \$902. The results were similar for LA.

5 EVALUATION

5.1 Balancing Fairness and Revenue

We now consider revenue maximization under the constraint of individual fairness. In Sec. 2 we referenced how this could be achieved after the choice of a distance function between outcomes, a distance function between users, and a linear objective function. Our choice of D, the distance between distributions of ads, is $D_{TV}(P,Q) = \frac{1}{2}\sum_{a \in A} |P(a) - Q(a)|$. For our choice of d, the distance score between users, we again use the distance of total variation, this time upon the histogram of visits to locations between each pair of users using the representation of users defined in Sec. 4.1. Our objective function is to maximize expected revenue, as defined as $\sum_{x \in V} g \cdot \mu_x(0) + t \cdot \mu_x(1)$ with g, the revenue of a generic ad, set to 1 and t, the revenue of a targeted ad set, to 2. After these choices, the linear program chooses a probability of shwoing a targeted ad to a user to maximize revenue under the constraints of similar users seeing similar ads.

In order to make the trade-off between revenue and fairness more fluid, we differ from prior work and introduce a new parameter k into Eq. 1:

$$D_{TV}(M(x), M(y)) \leq k \cdot d(x,y) \tag{3}$$

A large k means more flexibility in ad assignment but less individual fairness; $k = \infty$ means identical users can see completely different ads. In contrast, a low value of k constrains the problem more, with $k = 0$ meaning all users must have the same ad distribution.

We run this linear program for both cities at all granularity levels and for multiple choices of k. We then compute a real revenue with the function $\sum_{x \in V} g \cdot \mu_x(0) + t \cdot \mu_x(1) \cdot \mathbf{1}_{x \in I}$ with the set I denoting users who actually posted the target tag. Due to the number of constraints growing quadratically with the number of users, Here we are only able to present results for fairness by race and leave detailed analysis of gender for later work.

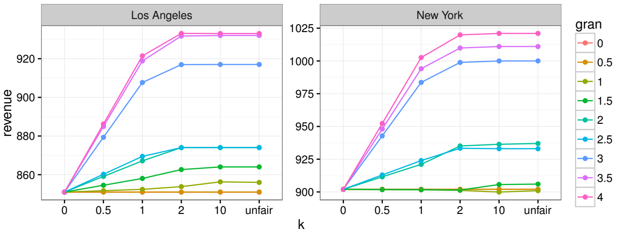

Figure 1: The impact of k and granularity impact on revenue.

Fig. 1 displays the impact of k and granularity on revenue for both cities with the tag fashion. The x axis corresponds to the choice of k used in the linear program. The y axis represents the actual

[2] https://www.facebook.com/business

revenue of the ad assignments output by the LP. Color denotes the granularity of location. The graph demonstrates again how finer granularity can increase revenue. In both NYC and LA, at nearly all values of k, a higher granularity corresponds to higher revenue. Another important takeaway is the shape of the lines. The revenue at $k = 2$ is nearly identical to the revenue at all higher amounts of k. The revenue declines rapidly at $k = 0$, where all individuals have the same distribution, and $k = 0.5$. The increase in revenue from $k = 1$ to higher values of k is significant but not a large portion of the highest optimal revenue, suggesting a good potential value due to its balance and simplicity.

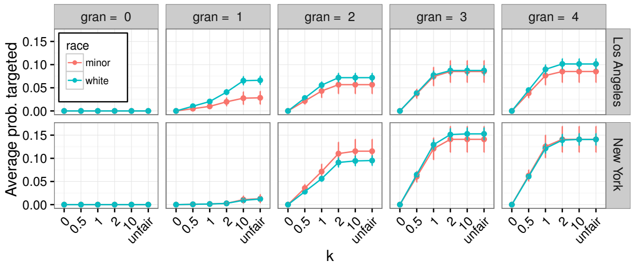

Figure 2: The impact of k and granularity on fairness.

We next examine the impact of k and granularity upon fairness. In Fig. 1, the x axis again corresponds to value of k. Color corresponds to race, with blue associated with caucasians and red associated with minorities. The y axis now corresponds to the average probability that users of the class saw a targeted ad, with error bars corresponding to standard error of the mean. Each facet represents a different level of granularity.

At lower levels of granularity, all users have similar low-resolution representations and thus it is difficult for our click predictor and then LP to risk displaying targeted ads, instead showing generic ads at all values of k. At medium level granularities, we see the algorithm begin to assign the ad to a small number of users and additionally the lines for each class to diverge, signally a rising level of group unfairness. Interestingly, in both graphs, the lines converge to be near identical at finer levels of granularity, at 4 digits for NYC and 3 and 3.5 digits for LA. This could be caused by mid-range granularities being associated more with neighborhoods, whereas very fine granularities will correspond to more exact venues, removing rougher associations of neighborhoods around areas with certain tags and narrowing them down to more specific places (e.g. 2 lat-long digits corresponds to roughly 1km, 4 to 10m).

5.2 Bounding Fairness

For two demographic attributes, race and gender, we compute the Earth Mover's Distance, using the pyemd package [6, 7]. More precisely, for race we calculate the EMD between two probability distributions, one over Caucasian users and the other over Non-Caucasian users, with the "distance" between users defined as the distance of total variation of the histogram of their location visits. Similarly, for gender we calculate the EMD between the distribution of female and male users. As mentioned in Section 2 we represented locations as "cells", assigning a photograph to a cell by truncating the latitude-longitude coordinates by a varying amount.

The large number of users labeled with gender presented a difficulty for our EMD calculation as Earth Mover's Distance does not scale well. We use agglomerative clustering [9] to approximate EMD. We found this technique that groups individuals into "points" is well suited to our problem due to nonuniform cluster sizes.

We add a mechanism to cope with statistical parity, as it may create a spurious statistical bias between finite size groups, even when the expectations among those groups are equal. In addition to computing EMD between demographic groups, we also computed EMD between randomly created groups with the same size as our demographic groups.

In Fig. 3 we show the result of this process. The x-axis shows the granularity in terms of latitude longitude decimal places. The y axis shows the EMD. Lines are colored according to demographic, and a dashed line indicates random grouping of users as opposed to grouping by demographics. To put the EMD numbers into perspective, on the lower end, an EMD of 0.05 means one group may be seeing a targeted ad 5% more often. At the higher end of 0.8, users across the two groups are seeing quite different sets of ads.

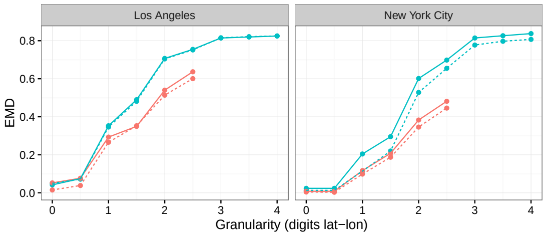

Figure 3: Risk vs. Granularity

In New York for race, the random line is clearly below the regular line, providing some evidence of real differences between the demographic groups as opposed to an artifact of sparsity. The line for gender is additionally more separate than it's counter-part in Los Angeles. This is possibly due to the much higher density in New York. As all users begin to have high difference scores from one another, caused by having no overlapping locations due to low density, all label assignments will be indistinguishable from each other. Gender overall seems to show a weaker separation between the real EMD and the random EMD.

The EMD increases as the data becomes more precise. One limitation of this study is that the distance d we chose does not distinguish two users who have nearby but non-intersecting visits and users who are on the opposite side of the city. Different choices of d with true geographical distance may refine those results.

6 CONCLUSION

In this work, we showed the impact of granularity on ad targeting, demonstrated the impact of fairness algorithms on a real world behavioral dataset, and explored a utility-fairness trade-off. There are many possible future directions. All results should be reproduced on larger datasets and different classes. One idea is to reformulate the problem in terms of *where* ads are shown or how users are reached, as opposed to focusing on the individuals. Building on our results, we also hope to create scalable algorithms for debiasing representations of users that work with sparse, large behavioral datasets.

REFERENCES

[1] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems (NIPS)*, July 2016.

[2] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Proceedings of Workshop FATML*, stat.AP, October 2016.

[3] Cynthia Dwork, M Hardt, T Pitassi, and O Reingold. Fairness through awareness. In *ITCS '12 Proceedings of the 3rd conference on Innovations in Theoretical Computer Science*, 2012.

[4] Alan Mislove, S Lehmann, Y Y Ahn, and J-P Onnela. Understanding the Demographics of Twitter Users. *ICWSM*, 2011.

[5] S Nilizadeh, A Groggel, P Lista, and S Das. Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility. *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*, 2016.

[6] O Pele and M Werman. Fast and robust earth mover's distances. *2009 IEEE 12th International Conference on Computer Vision*, 2009.

[7] Ofir Pele and Michael Werman. A linear time histogram metric for improved SIFT matching. In *Proceeding ECCV '08 Proceedings of the 10th European Conference on Computer Vision*, pages 495–508. Hebrew University of Jerusalem, Jerusalem, Israel, December 2008.

[8] Christopher J Riederer, Sebastian Zimmeck, Coralie Phanord, Augustin Chaintreau, and Steven M Bellovin. I don't have a photograph, but you can have my footprints.: Revealing the demographics of location data. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, pages 185–195. ACM, 2015.

[9] Joe H Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963.

[10] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW '17 Proceedings of the 26th International Conference on World Wide Web*, October 2016.

[11] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable Classification Models for Recidivism Prediction. *FATML*, March 2015.