Defining and Measuring Fairness in Location Recommendations

Leonard Weydemann
Travel Data Solution
Austria
leonard.weydemann@travel-data-solution.com

Dimitris Sacharidis
TU Wien
E-Commerce Research Unit
Austria
dimitris@ec.tuwien.ac.at

Hannes Werthner
TU Wien
E-Commerce Research Unit
Austria
hannes@ec.tuwien.ac.at

ABSTRACT

Location-based recommender systems learn from historical movement traces of users in order to make recommendations for places to visit, events to attend, itineraries to follow. As with other systems assisting humans in their decisions, there is an increasing need to scrutinize the implications of algorithmically made location recommendations. The challenge is that one can define different fairness concerns, as both users and locations may be subjects of unfair treatment. In this work, we propose a comprehensive framework that allows the expression of various fairness aspects, and quantify the degree to which the system is acting justly. In a case study, we focus on three fairness aspects, and investigate several types of location-based recommenders in terms of their ability to be fair under the studied aspects.

CCS CONCEPTS

• **Information systems → Location based services**; **Recommender systems**.

KEYWORDS

Fairness; Discrimination; Bias; Recommender Systems; Location-Based Recommendations

ACM Reference Format:
Leonard Weydemann, Dimitris Sacharidis, and Hannes Werthner. 2019. Defining and Measuring Fairness in Location Recommendations. In *3rd ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks, and Geoadvertising (LocalRec'19), November 5, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3356994.3365497

1 INTRODUCTION

More and more often nowadays machines make decisions on behalf of humans, or assist them in their choices, across an increasingly broad range of fields. These range from innocuous tasks, such as what to buy in a store, to more critical decisions, such as whether to grant a loan. It is thus natural to question the trustworthiness of such algorithmic decisions, and worry about their implications. An important issue to consider is fairness. Broadly speaking, fairness in machine-made decisions means that the machine should not discriminate against individuals. There is a long line of recent research concerning fairness in the area of machine learning [8, 22, 35, 37], information retrieval [1, 4, 28, 33, 36], and recommender systems [2, 3, 13, 17, 18, 21, 26, 29, 34].

Our focus is on algorithmic decisions made by *Location Recommenders* (LRs), where we study the degree to which they may cause unfairness. It is reasonable to question whether LRs discriminate against the users they are supposed to serve. For example, does an LR give recommendations of lower utility to individuals from a specific race group? Moreover, LRs have the potential to introduce another type of unfairness that concerns the objects to be recommended, i.e., places such as points/areas of interest, or sequences thereof (e.g., routes, itineraries). In LRs, places compete against each other for reaching the attention of the users. It is thus equally important to investigate whether an LR promotes a healthy competition for these places. For example, is a specific small business being exposed less than a corresponding large enterprise?

In this work, we propose a simple, but powerful framework that enables the formulation of various fairness criteria, as the aforementioned, that arise in LRs. Existing literature on recommendation fairness is concerned with either ensuring user- or item- fairness, e.g., as defined in [2], and fails to systematically study the area. To the best of our knowledge, this is the first work that investigates fairness in the domain of LRs.

To investigate a fairness criterion, the proposed framework requires the specification of two components. The first is the *probe* probability distribution that captures how the LR makes recommendations over a period of time. For example, we may want to observe the probability with which locations are recommended to some sensitive race group (e.g., a minority), also called the protected group. Thus, by observing the LBR recommendations to the protected group members over some time, we can empirically calculate this probe distribution. The second component is the *target* probability distribution that prescribes how an ideal, perfectly fair LR should behave. Continuing the example, we may want the locations to be recommended with equal probability to the protected group as to the non-protected. As before, by observing the LR recommendations to the non-protected users, we empirically calculate the target distribution. We can then quantify the degree to which the LBR is being fair, according to the desired fairness criterion, by measuring how far from the target the probe distribution is.

We illustrate the expressiveness of our framework by formulating three criteria that convey distinct fairness concerns: the popularity biases of locations should not be amplified, recommendations are agnostic to user nationalities, and recommendations should respect prior user preferences. We then investigate them in the context of a case study with real traces of users' movement in the city of Vienna.

LocalRec'19, November 5, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6963-3/19/11…$15.00
https://doi.org/10.1145/3356994.3365497

Specifically, we implement five types of location recommenders, and employ them to synthetically create recommendations for the users based on their history and their current locations. We log these recommendations over time, appropriately define probe and target distributions for each fairness criterion, and ultimately quantify the fairness of each LR. The objective of this case study is on the one hand to demonstrate our framework, and on the other hand to investigate which recommender type is more fair.

The remaining of this paper is structured as follows. Section 2 overviews methods for recommending locations, and fairness definitions and approaches for recommendations. Section 3 presents our fairness framework for LRs, and introduces the three criteria we study. Then, Section 4 discusses the case study and its findings. Finally, Section 5 concludes this paper and sketches future work.

2 RELATED WORK

Location Recommenders. We briefly overview methods that provide location recommendations, e.g., where to go next. In the simplest approach, locations can be treated as items in conventional recommenders, and hence standard *collaborative filtering* (CF) techniques apply. The most well-known of these is *matrix factorization* (MF), e.g., [14, 25] that learns a latent space in which users and items are embedded, so as the distance (in terms of the inner product) between a user and an item defines the degree of match between the two. While basic MF is agnostic of spatial relationships, some methods, e.g., [5], also model geographical influence. More recent CF techniques, e.g., [32], employ neural-network-based *autoencoders* (AE) to derive embeddings for users and items.

Transitions between locations convey important information about users' preferences that should be exploited by an LR. Therefore, most recent LR techniques employ sequence-aware recommendation techniques [23]. The simplest of them is to use *Markov chains* (MC), where the empirical transition probability between two locations (or two location classes) is used to predict the next location given the current, e.g., as in [19, 38]. The use of factorization techniques to define personalized Markov chains has also been proposed [15, 24]. Recurrent neural networks can also be employed to learn from the transitions among locations, e.g., as in [9]; however their effectiveness compared to conventional CF has been questioned [11].

Another approach to learn from sequences of locations is to apply techniques from natural language processing, treating thus locations as words and trajectories as sentences. For example, the well-known word2vec method [20], can be used to create a *contextual embedding* (CE) of a location based on its context (i.e., the locations shortly before and after it in trajectories), similar to [7]. Moreover, metadata about the locations could also be utilized to construct *metadata contextual embeddings* (MCE), similar to [31].

Recommendation Fairness. Fairness, as discussed in [2], may concern the users that receive recommendations (i.e., the consumers, hence C-fairness), or the items that are being recommended (i.e., the providers/producers behind the items, hence P-fairness); in some cases, fairness may concern both stakeholders.

In user-fairness, users may belong to *protected groups*, e.g., based on their race or gender, and the broad goal is to treat these groups equally. In one line of work [13], the focus is on *demographic* (a.k.a. statistical or group) *parity*, where the predicted ratings of users across groups for the same item should match. This strong condition can be relaxed to the notion of *equal opportunity* [8], where, in the context of recommenders, it means that prediction errors should occur with the same frequency and magnitude across the groups [34]. [3] considers top-N recommendations, and counts the number of recommended items from some desired class of items, e.g., a chosen genre. Fairness is achieved when the counts are equal among the protected and non-protected users. The proposed approach seeks to balance the user-neighborhood formation in collaborative filtering, so that it assigns equal total weight to members from the two user groups. A related goal is to achieve fairness when making recommendations to a group of users, i.e., so as not to displease any individual [17, 26].

In item-fairness, there may be various *protected classes* of items, e.g., items provided by a particular group of people, and the broad goal is to not discriminate against these classes. In [18], fairness is achieved when a recommendation list covers all classes, i.e., it contains at least one item from every (protected) class. A different variant of item fairness is considered in [29, 30]. The goal is to provide recommendations that have similar class distribution with the user's observed preferences, e.g., if a user has watched 7 romance and 3 action movies, she should get recommended about 70% romance and 30% action movies.

Our proposed fairness framework for location-based recommenders is flexible enough to accommodate most of the aforementioned fairness aspects. Compared to previous work, our framework is not restricted into the binary user-/item- fairness classification prevalent in the recommender systems literature, and thus enables formulating and quantifying additional, interesting fairness aspects.

3 FAIRNESS IN LOCATION RECOMMENDATIONS

We consider a location recommender system (LR). In what follows, we assume that all *locations* belong to a predefined set $\mathcal{L}$ of points/areas of interest (P/AoI). Let $u \in \mathcal{U}$ denote an LR user, and let H_u denote her *history*, represented as a set of trajectories, i.e., sequences of timestamped locations (t_i, ℓ_i), where $\ell_i \in \mathcal{L}$. Given a set of users and their histories, the goal of the LR is to recommend to a target user possible locations to visit next. For a target user u at current location ℓ_u, let R_u denote the *recommendation*, i.e., a ranked list of locations, outputted by the LR.

Observation Matrix. Consider a set of recommendations made by the LR over a period of time, concerning potentially multiple users. We define the observations O of the LR, as the matrix where entry $O[u, \ell]$ counts the number of times location ℓ was recommended to user u in the period. Note that although we ignore it, it is possible for the observations matrix to account for the positional bias [12] in ranked lists — a location recommended at the top of the list is exposed more to the user than another (or the same) location ranked fifth.

The observations matrix defines an empirical distribution over user-location pairs. Specifically, the probability that the LR may recommend location ℓ to user u is $P(u, \ell) \approx \frac{O[u,\ell]}{\sum_{u'} \sum_{\ell'} O[u',\ell']}$. (In case positional bias is accounted for, we may say that $P(u, \ell)$ is the probability with which u is exposed to ℓ.)

In the following, we first discuss how OLAP-like operations roll-up and slice, [6], can be applied to the observations matrix so as to define other empirical distributions of interest. Then, we specify the two components necessary to define various fairness criteria, and finally discuss three of them.

Roll-Up. Users are associated with demographic factors that partition them into *groups* $\mathcal{G}$. For example, if we assume gender and race as the factors of interest, a demographic group is black females. Often, a certain group is identified as *protected*, and the goal is to investigate whether discrimination against its members occurs. As individuals may not fall into a single group (e.g., a mixed-race person), we define a membership probability $p(g|u)$ to indicate the degree user u is a member of group g; naturally, $\sum_{g \in \mathcal{G}} p(g|u) = 1$.

Similarly, locations are described by some attributes that assign them into *classes* $\mathcal{K}$. We assume that a location ℓ may belong to a class $\kappa \in \mathcal{K}$ with some membership probability $p(\kappa|\ell)$; for any location ℓ, it holds that $\sum_{\kappa \in \mathcal{K}} p(\kappa|\ell) = 1$. Location classes could represent PoI categories, so that a location representing a mall, for example, can be a member of the shopping and the entertainment classes. Alternatively, classes could correspond to types of ownership, so that a location can be classified as a small, medium, or large enterprise, or could even correspond to business owners.

Groups and classes represent hierarchies for users and locations, respectively. The observations matrix can be *rolled-up* along these hierarchies. For a group g and location ℓ, the corresponding entry of the rolled-up observations matrix is computed as $O[g, \ell] = \sum_u p(g|u)O[u, \ell]$. Similarly, we compute other roll-ups as $O[u, \kappa] = \sum_\ell p(\kappa|\ell)O[u, \ell]$, and $O[g, \kappa] = \sum_u \sum_\ell p(g|u)p(\kappa|\ell)O[u, \ell]$. Moreover, we can roll users or locations up to the top of the hierarchy. For example, keeping locations at the highest granularity and rolling users up to the top, we obtain $O[\ell] = \sum_g \sum_u p(g|u)O[u, \ell]$.

Interpreting the rolled-up observations matrix as an empirical distribution, we may reach similar conclusions as before. For example, the probability with which a location class κ is recommended to a user group g is approximately $P(g, \kappa) \approx \frac{O[u,\kappa]}{\sum_{u'} \sum_{\kappa'} O[u',\kappa']}$. Moreover, marginal probabilities for some location/class or user/group can be expressed in terms of the observations matrix. For example, the probability with which a location is recommended is $P(\ell) \approx \frac{O[\ell]}{\sum_{\ell'} O[\ell']}$.

Slice. Given a (rolled-up) observations matrix, we may potentially focus on a specific user (or group), or an item (or class). In this case, we *slice* the matrix and the observations reduce to a vector. For example, we may want to focus on a particular minority group, and observe what locations are being recommended by the LR.

Taking a slice provides an estimate of the conditional probability of a user/group subject to a location/class pair, or vice versa. For example, the probability of recommending location ℓ to a protected group g is approximately $P(\ell|g) \approx \frac{O[g,\ell]}{\sum_{\ell'} O[g,\ell']}$.

Fairness. Different concepts of fairness can be operationalized by specifying two components. The first is the *probe distribution* Q, which can be: a joint user/group – location/class probability; a marginal user/group or location/class probability; or a conditional on user/group or location/class probability. The probe Q is approximated from the observations matrix after performing appropriate roll-up and slice operations. For example, we may want to express a fairness concept that concerns what locations are recommended to a specific group g of users. In this case, the probe distribution is $Q = P(\ell|g)$, estimated from a slice of the observations matrix after rolling users up to groups.

The second component is the *target distribution* T that specifies the ideal probe distribution, which captures the condition of perfect fairness. Continuing the previous example, we define perfect fairness as the case when each location is equally likely to be recommended to group g, i.e., T is the uniform distribution over locations.

In the ideal case when the probe Q and target T distributions coincide, we say that the LR is *perfectly fair*. In general, however, we wish to quantify the possible deviation of the probe distribution from the target. Dissimilarity between two distributions can be captured by divergence measures, such as the Kullback–Leibler divergence (KL-div) [16]. The KL-Div of Q from T is defined as:

$$D(Q \| T) = \sum_x T[x] \cdot \log \left(\frac{T[x]}{\tilde{Q}[x]} \right),$$

where $\tilde{Q} = (1 - \alpha) \cdot Q + \alpha \cdot T$ is a smoothed version of the probe distribution (using some small constant α, e.g., 0.001) that eliminates non-defined values for KL-div.

The closer Q to T is, the lower the KL-div is, hence the more fair the LR is. When Q and T coincide, KL-div becomes zero.

Fairness Criteria. The aforementioned framework is able to capture various fairness criteria that might be of interest in location-based recommenders. In the following, we present three distinct fairness criteria, which we further explore in our case study. The list is only indicative of the formulations possible, and is in no way exhaustive.

F1. Suppose we want to observe whether the LR is amplifying existing popularity biases of locations. In this sense, an LR is seen as fair if it does not unjustly promote some locations at the expense of others. We consider as probe distribution that of the marginal probability $P(\ell)$ of a location being recommended; this is estimated from the observation matrix by rolling up all users. As the target, we calculate a prior base distribution of location popularity from some historical traces. Given these definitions, an LR is considered fair if, over some sufficiently long time period, it respects the prior popularity of locations.

F2. We want to investigate whether the LR makes recommendations that are agnostic to user nationalities. We consider as probe the conditional probability $P(\ell|g)$ that a location ℓ is recommended to a particular nationality group g; the probe is estimated from the observations by rolling users up to groups and slicing for g. As the target, we set it to the marginal probability $P(\ell)$ of a location being recommended (the probe used in F1). In this scenario, an LR is considered fair when locations are recommended to nationality g with the same probability as for all users, i.e., the conditional $P(\ell|g)$ matches the marginal $P(\ell)$.

F3. Another fairness criterion is whether the LR respects the prior users' preferences, in terms of location classes. For example, if a user group tends to visit 40% bars, 40% restaurants, and 20%

theaters in her past, the system should recommend locations with the same class mix, i.e., it should be calibrated [29]. To explore F3, we set as probe the conditional probability $P(\kappa|g)$ that a location class κ will be recommended given group g; this is estimated from the observations by rolling locations up to classes, users up to groups, and then slicing for group g. For the target, we calculate the prior location class distribution of group g from its history. Then, the LR is fair for group g when it provides recommendations that match the group's historical preferences.

4 CASE STUDY

Section 4.1 describes the data used, the location recommenders tested, and the evaluation methodology. Section 4.2 presents an evaluation based on our fairness framework considering the three criteria defined in Section 3.

4.1 Setup

Data. The data used in this study is provided by Travel Data Solution, a company that equips rooms of selected hotels in Austria with cellular-based mobile hotspots. Hotel guests are allowed to take these portable devices outside of the hotel to enjoy free internet connectivity along their daily visits. In that case, the device is able to collect anonymized location data stemming from the activity of the guests, along with timestamps. Location coordinates are computed from the triangulation of the cell towers the device is connected to, and thus are quite noisy. The raw location data contains entries of the form device-id, timestamp, latitude, and longitude. Additionally, standard profile and demographics data is collected per device from a questionnaire presented at sign-up. For the purposes of this study, hotel guests/devices are the *users*, while their nationality is used to define *groups*; we focus on the three largest groups, USA, China, and Russia.

As raw coordinates come with considerable uncertainty, we coarsen them based on a square grid, which is centered at the city center (Stephansplatz), has a side length of 15km, and is partitioned into 64 by 64 square cells. Each cell has a side length of about 230m, and can be crossed by foot in 5 minutes. The grid cells correspond to the *locations* in our framework. The *classes* of these locations are compiled using Foursquare data. Specifically, for each cell we retrieve from Foursquare all venues that are within, along with their top-level categories. We then compute the probability distribution of categories in a cell, which defines the class memberships. There are 8 classes, corresponding to the Foursquare top-level categories: arts and entertainment, education, food, nightlife, outdoors, professional, shop, travel.

User trajectories are processed as follows. First, we remove coordinates that are outliers, using the metric proposed in [10]. We then split a trajectory into sub-trajectories, so that a sub-trajectory contains consecutive locations that are less than 25 minutes apart; we empirically found the 25 minute rule to result in trajectories that better match actual trips. Then we coarsen the trajectory based on the grid, so that it is a sequence of timestamped locations (grid cells), and extract *trips*. We are interested in trips done by foot so we only consider trips with a mean speed of 1.4 m/s. Moreover, we remove parts of the trips that contain transitions between non-adjacent locations. Finally, we retain trips that contain at least 5 unique locations. The cleaned data contains 1,418 trips made by 539 users during the summer of 2019 (2019-06-01 to 2019-08-31) in the city of Vienna.

Recommenders. We examine five different types of LRs. Given the current user location, each system recommends a location among the eight neighboring locations (grid cells). The LRs are trained and tuned using the historical trips. Specifically, the historical trips are partitioned into a training and a validation set, so that around 70% of each trip is contained in the training set. An LR is trained on the training set, while its hyperparameters are selected using the validation set. We defined a parameter grid to search for an optimal combination of hyperparameters. We varied the number of epochs between 50 and 200. The learning rate was selected from values ranging between 0.001 and 0.5, and for the weight-decay between 0.0001 and 0.1. For matrix factorization the latent space dimensionality was chosen between 10, 50, and 100. The same goes for the number of hidden layers in the autoencoder as well as the context embedding recommenders. The context size was varied between 2, 3 or 4; higher numbers were deemed unnecessary as the trip length rarely exceeds 8. We chose the parameters with the highest mean reciprocal rank.

MF. The matrix factorization LR operates on a ratings matrix, where a "rating" represents the number of times a user has visited the location. The latent space dimensionality is set to 10.

AE. In the autoencoder LR, a neural network with a single hidden layer of size 50 is used to encode and decode a user's trips. Specifically, the input is a vector of size 4,096 representing the number of times each location was visited by the user.

MC. In the Markov chain LR, transitions among locations are recorded in a transition matrix. Similar to [27], we employ skipping to reduce the sparseness of the matrix. To make a recommendation, the transitions from the current location to each neighboring locations are examined, and the most probable is returned.

CE. For the contextual embedding LR, we set the context size to 2, i.e., each location is associated with a context defined by its preceding and proceeding locations. Given its context, the neural network learns to predict the location. The embedding dimensionality is 100, and a single hidden linear layer of size 128 is used before the output.

MCE. The metadata contextual embedding LR differs from CE in that it also considers the class description of locations. Specifically, for a target location, the neural network takes as input the class descriptions of the target's context, and learns to predict the class description of the target. To make a recommendation, MCE returns the neighboring location that has the highest cosine similarity to the output class distribution. A hidden layer of size 128 is used. Context size was set to 4.

Observations. As none of the recommenders are actually deployed, we create a synthetic sequence of recommendations. For each trip, we first ask for a location recommendation based on the last location in the trip contained in the training set. The top recommendation is then assumed to be the next current location, and a subsequent recommendation is requested. We repeat this process for each trip from 10 up to 100 times, which determines the *observation length.*

4.2 Results

In what follows, we investigate how fair are the different LRs with respect to the three criteria defined in Section 3.

Fairness Criterion F1. In this criterion, we assume there is a prior distribution of popularity among locations, and want to observe whether an LR makes recommendations following this distribution or amplifies existing biases. We use the trips in the training set to compute the popularity distribution of locations, which forms the target as per our framework.

We first investigate F1 visually. Figure 1(a) depicts the popularity distribution as a heatmap over the city; the darker the color, the more popular a location is. This corresponds to the target distribution that an LR should adhere to in order to be considered fair according to F1. For each of tested LRs, the top (resp. bottom) row in Figure 1(b–f) presents as a heatmap the distribution of the recommended locations after observing 20 (resp. 100) recommendations per user.

The following observations are of interest. The various recommenders exhibit distinct heatmap patterns. For example, compared to the target, matrix factorization (MF) increases the popularity of neighboring locations. In contrast, autoencoders (AE), and contextual embeddings (CE and MCE) tend to emphasize certain popular locations. The Markov Chain (MC) method appears to produce distributions that vary the most when going from 20 to 100 recommendations per user. Overall, it appears that MF more closely adheres to the prior popularity distribution, with AE, CE, MCE producing more skewed distributions, and MC less skewed distributions.

We next investigate F1 quantitatively. For each LR, we consider the probe distribution at an observation length varying from 10 to 100, and compute its KL-Div with respect to the target distribution. In Figure 2, we draw the KL-Div values as a function of the observation length. A lower KL-Div values means that the probe distribution better matches the target distribution and thus the LR is considered more fair. Thus, among all recommenders, MF is being more fair with respect to the F1 fairness criterion. In contrast, MCE is less fair with a divergence score that slightly increases as more recommendations are observed. These findings are inline with the observations made from the heatmaps.

Fairness Criterion F2. In the second fairness criterion we investigate, a location recommender is considered fair when different user groups (nationalities) receive similar recommendations. The target distribution is the marginal probability with which locations are being recommended, irrespective of the user's group. The probe distribution is the conditional probability of locations being recommended to a group.

As before, we first investigate F2 visually. Figures 3, 4, and 5 depict the probe distributions as heatmaps for groups USA, China, and Russia, respectively. A recommender is fair, according to F2, when these heatmaps look similar across groups. Across the figures, MF appears to recommend similar places to the three groups. In contrast, MC and MCE recommendations induce distinctively different patterns for the groups.

The quantitative analysis of F2 verifies the aforementioned findings. Specifically, for each LR, we compute the KL-Div of the three probe distributions to the target, and report the results in Figure 6. Recommenders MF and MC have the lowest KL-Div, but by a small margin, and are thus considered the most fair under F2.

Fairness Criterion F3. In the last examined criterion, an LR is considered fair if for each group it respects the prior preferences expressed over location classes. In this case, the target distribution is the historical distribution over classes in the group, whereas the probe distribution is the probability with which a particular class is being recommended.

Figures 7, 8, and 9 represent the target and probe distributions of location classes for groups USA, China, and Russia, respectively. In general, all methods appear to perform well in terms of the F3 fairness aspect. For the first group, MF and AE appear to better match the target distribution, for the second group, CE, while for the third group, AE and MF.

These conclusions are also supported by Figure 10, which depicts the KL-Div between the target and the probe distributions for each method and group. Overall, the AE recommender is consistently the most fair according to F3.

Discussion. The primary goal of the case study is to demonstrate how seemingly different fairness concerns can be expressed in a unifying framework, and also quantified. The secondary goal is to explore how fair various types of location recommenders can be. While the three criteria we focus on are in no way exhaustive, we can draw some conclusions. Specifically, among the tested recommenders, latent factor models such as MF and AE appear to be consistently more fair than others.

5 CONCLUSION

In this work, we discussed fairness aspects that are relevant to location-based recommenders. We argued that there can be various, quite distinct interpretations of what a fair system means, and that there is lack of focus in the literature in such issues. We then propose a framework that not only allows the formulation of a wide range of fairness concepts, but also provide the means to quantify how fair a recommender is. Using this framework, we describe three distinct fairness criteria, and present a thorough case study, involving real mobility traces to investigate the degree of fairness achieved by five types of recommenders.

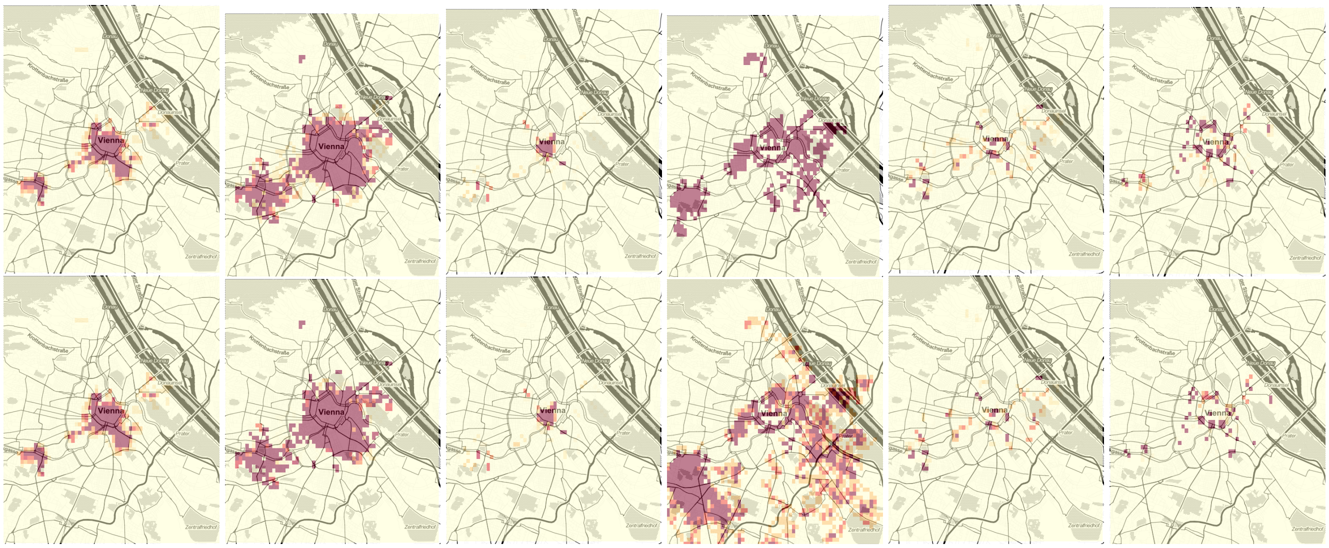
An apparent direction for future research is the design of fairness-aware location-based recommenders that aim to satisfy given fairness criteria as much as possible. As fairness often comes at the expense of utility, one must also study the associated trade-offs, and consider methods that seek to achieve a balance between utility and fairness.

ACKNOWLEDGMENTS

The authors would like to thank FLINK for supplying the data in collaboration with Travel Data Solution, and especially Marc Melich for making this collaboration possible.

REFERENCES

[1] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *ACM SIGIR*. ACM, 405–414. https://doi.org/10.1145/3209978.3210063
[2] Robin Burke. 2017. Multisided Fairness for Recommendation. *CoRR* abs/1707.00093 (2017). arXiv:1707.00093 http://arxiv.org/abs/1707.00093



(a) Target (b) MF (c) AE (d) MC (e) CE (f) MCE

Figure 1: Fairness Criterion F1: (a) Location heatmap for the target distribution. (b–f) Location heatmaps for the probe distributions of the five LRs. Top (resp. bottom) row is for an observation length of 20 (resp. 100) recommendations per user.

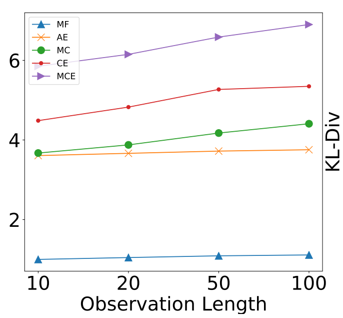


Figure 2: Fairness Criterion F1: KL-Div of the probe location distribution induced by each LR with respect to the target distribution of prior location popularity.

[3] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research)*, Vol. 81. PMLR, 202–214. http://proceedings.mlr.press/v81/burke18a.html

[4] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *ICALP (LIPIcs)*, Vol. 107. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 28:1–28:15. https://doi.org/10.4230/LIPIcs.ICALP.2018.28

[5] Chen Cheng, Haiqin Yang, Irwin King, and Michael R. Lyu. 2012. Fused Matrix Factorization with Geographical and Social Influence in Location-Based Social Networks. In *AAAI*. AAAI Press.

[6] Jim Gray, Adam Bosworth, Andrew Layman, and Hamid Pirahesh. 1996. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total. In *ICDE*. IEEE Computer Society, 152–159.

[7] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. E-commerce in Your Inbox: Product Recommendations at Scale. In *KDD*. ACM, 1809–1818.

[8] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*. 3315–3323. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning

[9] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent Neural Networks with Top-k Gains for Session-based Recommendations. In *CIKM*. ACM, 843–852.

[10] Boris Iglewicz and David Hoaglin. 1993. Volume 16: how to detect and handle outliers. *The ASQC basic references in quality control: statistical techniques* 16 (1993).

[11] Dietmar Jannach and Malte Ludewig. 2017. When Recurrent Neural Networks meet the Neighborhood for Session-Based Recommendation. In *RecSys*. ACM, 306–310.

[12] Thorsten Joachims and Filip Radlinski. 2007. Search Engines that Learn from Implicit Feedback. *IEEE Computer* 40, 8 (2007), 34–40. https://doi.org/10.1109/MC.2007.289

[13] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2013. Efficiency Improvement of Neutrality-Enhanced Recommendation. In *RecSys Workshops (CEUR Workshop Proceedings)*, Vol. 1050. CEUR-WS.org, 1–8. http://ceur-ws.org/Vol-1050/paper1.pdf

[14] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *SIGKDD*. 426–434. https://doi.org/10.1145/1401890.1401944

[15] Defu Lian, Vincent Wenchen Zheng, and Xing Xie. 2013. Collaborative filtering meets next check-in location prediction. In *WWW (Companion Volume)*. International World Wide Web Conferences Steering Committee / ACM, 231–232.

[16] F. Liese and Igor Vajda. 2006. On Divergences and Informations in Statistics and Information Theory. *IEEE Trans. Information Theory* 52, 10 (2006), 4394–4412. https://doi.org/10.1109/TIT.2006.881731

[17] Xiao Lin, Min Zhang, Yongfeng Zhang, Zhaoquan Gu, Yiqun Liu, and Shaoping Ma. 2017. Fairness-Aware Group Recommendation with Pareto-Efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, August 27-31, 2017*. ACM, 107–115. https://doi.org/10.1145/3109859.3109887

[18] Weiwen Liu and Robin Burke. 2018. Personalizing Fairness-aware Re-ranking. *CoRR* abs/1809.02921 (2018). arXiv:1809.02921 http://arxiv.org/abs/1809.02921

[19] Xin Liu, Yong Liu, Karl Aberer, and Chunyan Miao. 2013. Personalized point-of-interest recommendation by mining users' preference transition. In *CIKM*. ACM, 733–738.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 3111–3119.

[21] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren G. Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *WWW*. ACM, 677–686. https://doi.org/10.1145/2566486.2568012

[22] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*. ACM, 560–568. https://doi.org/10.1145/1401890.1401959

(a) MF (b) AE (c) MC (d) CE (e) MCE

Figure 3: Fairness Criterion F2, Group USA: Location heatmaps for the probe distributions of the five LRs.

(a) MF (b) AE (c) MC (d) CE (e) MCE

Figure 4: Fairness Criterion F2, Group China: Location heatmaps for the probe distributions of the five LRs.

(a) MF (b) AE (c) MC (d) CE (e) MCE

Figure 5: Fairness Criterion F2, Group Russia: Location heatmaps for the probe distributions of the five LRs.

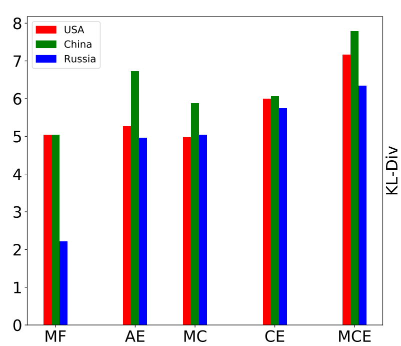


Figure 6: Fairness Criterion F2: KL-Div of the probe location distribution induced by each LR for each group with respect to the target location distribution that ignores groups.

[23] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-Aware Recommender Systems. *ACM Comput. Surv.* 51, 4 (2018), 66:1–66:36.

[24] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *WWW*. ACM, 811–820.

[25] Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic Matrix Factorization. In *NIPS*. 1257–1264. http://papers.nips.cc/paper/3208-probabilistic-matrix-factorization

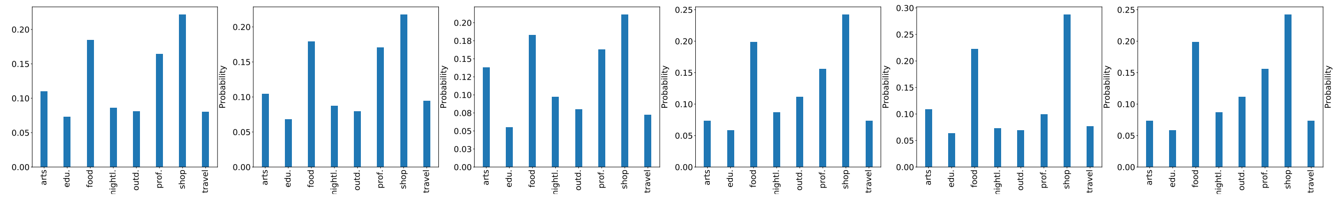
[26] Dimitris Serbos, Shuyao Qi, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2017. Fairness in Package-to-Group Recommendations. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. ACM, 371–379. https://doi.org/10.1145/3038912.3052612

[27] Guy Shani, Ronen I. Brafman, and David Heckerman. 2002. An MDP-based Recommender System. In *UAI*. Morgan Kaufmann, 453–460.

[28] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *ACM KDD*. ACM, 2219–2228. https://doi.org/10.1145/3219819.3220088

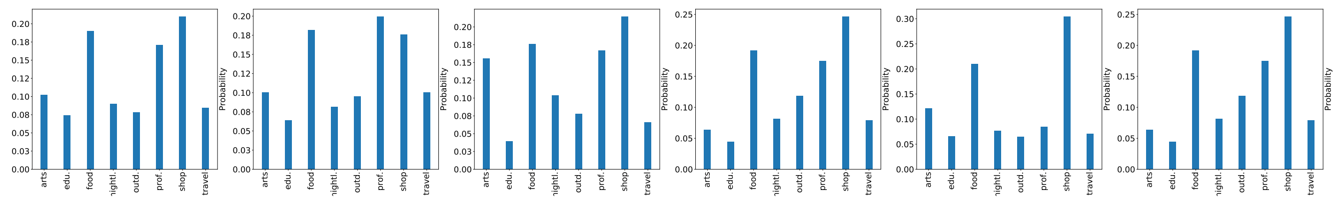
[29] Harald Steck. 2018. Calibrated recommendations. In *ACM RecSys*. ACM, 154–162. https://doi.org/10.1145/3240323.3240372

[30] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. 2018. Bias Disparity in Recommendation Systems. *CoRR* abs/1811.01461 (2018). arXiv:1811.01461 http://arxiv.org/abs/1811.01461



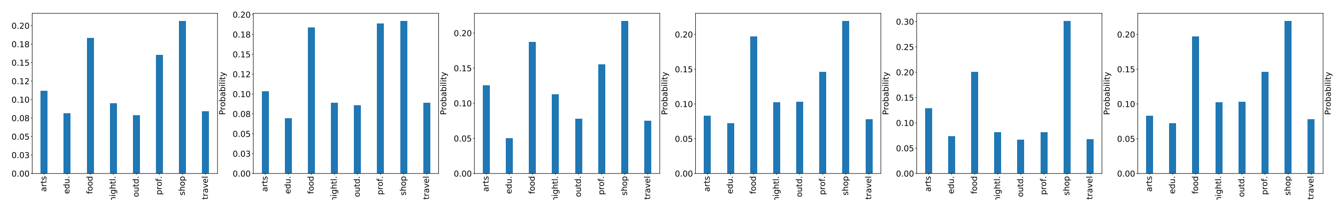
(a) Target (b) MF (c) AE (d) MC (e) CE (f) MCE

Figure 7: Fairness Criterion F3, Group USA: (a) Item classes histogram for the target distribution. (b–f) Item classes histograms for the probe distributions of the five LRs.



(a) Target (b) MF (c) AE (d) MC (e) CE (f) MCE

Figure 8: Fairness Criterion F3, Group China: (a) Item classes histogram for the target distribution. (b–f) Item classes histograms for the probe distributions of the five LRs.



(a) Target (b) MF (c) AE (d) MC (e) CE (f) MCE

Figure 9: Fairness Criterion F3, Group Russia: (a) Item classes histogram for the target distribution. (b–f) Item classes histograms for the probe distributions of the five LRs.

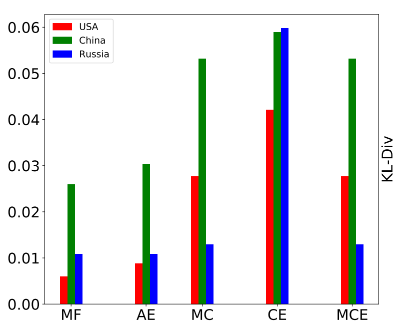


Figure 10: Fairness Criterion F3

[31] Flavian Vasile, Elena Smirnova, and Alexis Conneau. 2016. Meta-Prod2Vec: Product Embeddings Using Side-Information for Recommendation. In *RecSys*. ACM, 225–232.

[32] Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. 2016. Collaborative Denoising Auto-Encoders for Top-N Recommender Systems. In *WSDM*. ACM, 153–162.

[33] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*. ACM, 22:1–22:6. https://doi.org/10.1145/3085504.3085526

[34] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *NIPS*. 2925–2934. http://papers.nips.cc/paper/6885-beyond-parity-fairness-objectives-for-collaborative-filtering

[35] 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 *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. ACM, 1171–1180. https://doi.org/10.1145/3038912.3052660

[36] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo A. Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *ACM CIKM*. ACM, 1569–1578. https://doi.org/10.1145/3132847.3132938

[37] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013 (JMLR Workshop and Conference Proceedings)*, Vol. 28. JMLR.org, 325–333. http://jmlr.org/proceedings/papers/v28/zemel13.html

[38] Jia-Dong Zhang, Chi-Yin Chow, and Yanhua Li. 2014. LORE: exploiting sequential influence for location recommendations. In *SIGSPATIAL/GIS*. ACM, 103–112.