FARGO: a Fair, context-AwaRe, Group recOmmender system

Davide Azzalini[1]*, Elisa Quintarelli[2], Emanuele Rabosio[1], and Letizia Tanca[1]

[1] Politecnico di Milano, Milan, Italy
[2] University of Verona, Verona, Italy
{davide.azzalini,emanuele.rabosio,letizia.tanca}@polimi.it
elisa.quintarelli@univr.it

Abstract. Lots of activities, like watching a movie or going to the restaurant, are intrinsically group-based. To recommend such activities to groups, traditional single-user recommendation techniques cannot be adopted, as a consequence, over the years, a number of group recommender systems have been developed. Recommending to groups items to be enjoyed together poses many ethical challenges, in fact, a system whose unique objective is to achieve the best recommendation accuracy possible, might learn to disadvantage submissive users in favor of more aggressive ones. In this work we investigate the ethical challenges of context-aware group recommendations, in the more general case of ephemeral groups (i.e., groups where the members might be together for the first time), using a method that can recommend also items that are new in the system. We show the goodness of our method on two real-world datasets. The first one is a very large dataset containing the personal and group choices regarding TV programs of 7,921 users w.r.t. sixteen contexts of viewing. The second one, which has been collected specifically for this work and that is made publicly available as one of the contributions of this article, gathers the musical preferences (both individual and in groups) of 280 real users w.r.t. two contexts of listening. We compare the results of our approach with seven other group recommender systems specifically developed to be fair. We evaluate the goodness of our recommendations using recall, while their fairness is assessed using two measures found in the literature, namely, score disparity and recommendation disparity. Our extensive experiments show that our method always manages to obtain the highest recall while delivering ethical guarantees in line with the other fair group recommender systems tested.

Keywords: group recommender systems · context-aware recommender systems · computer ethics · fairness.

* Corresponding author.
DA is supported by the ABB-Politecnico di Milano Joint Research Center, which provided financial support.

1 Introduction

Several everyday activities are intrinsically group-based, thus recent research concentrates also on systems that suggest activities that can be performed together with other people and are typically social. The group recommendation problem introduces further challenges with respect to the traditional single-user recommendations: *(i)* the group members may have different preferences, and finding items that meet the tastes of everyone may be impossible; *(ii)* a group may be formed by people who happen to be together for the first time, and, in this case, not being available any history of the group's preferences, the recommendation can only be computed on the basis of those known for the group members combined by means of some aggregation function; *(iii)* last but not least, people, when in a group, may exhibit different behaviors with respect to when they are alone, and therefore their individual preferences sometimes might not be a reliable source of information.

This last observation introduces an unfairness problem, in fact, if the recommender system learns to consider the preferences of some users as more relevant than those of the others, the overall satisfaction of the users belonging to a group may not be optimal. This unbalance in the negotiation power that the system learns to assign to different users with the purpose of obtaining the best possible recommendation accuracy may be the result of unfair dynamics, such as some users being more aggressive and some others not feeling confident enough to stand up for themselves.

In this work we extend a state-of-the-art system for context-aware recommendations to ephemeral groups based on the concept of contextual influence [1, 2] to account also for fairness.

Experiments on two real-world datasets show that the proposed approach outperforms seven other fair group recommender systems by achieving a consistently better recall while providing similar ethical guarantees.

The main original contributions of this article are: *(i)* a novel technique for providing fair, context-aware recommendations to ephemeral groups able to recommend also items that are new in the system; *(ii)* an extensive experimental campaign on two real-world datasets which demonstrates the goodness of our technique both in terms of accuracy and fairness of the recommendations produced; *(iii)* the first publicly available real-world dataset with both individual and group context-aware preferences. The dataset can be downloaded at `https://github.com/azzada/FARGO`.

2 Related Work

In this section, works related to the one presented in this paper are reviewed. After a brief general introduction to recommender systems, context-awareness, and group recommendation techniques, a thorough discussion of fairness in group recommender systems is presented.

Recommender Systems are software tools and techniques providing suggestions for items to be of use to a user [3].

Context-aware Recommender Systems. The majority of the existing approaches to Recommender Systems do not take into consideration any contextual information, however, in many applications, it may not be sufficient to consider only users and items [4]. Recent studies have shown that Context-Aware Recommender Systems can generate a very high increase in performance [5].

Group Recommender Systems. Group Recommender Systems are systems which produce a recommendation for a group of users [6]. Group recommendations works usually address two kinds of groups, *persistent* and *ephemeral* [7]. Persistent groups have a previous significant history of activities together, while ephemeral groups are formed by people who may happen to be together for the first time. In the case of persistent groups, classical recommendation techniques can be used since the group can be considered as a single user; whereas in the case of ephemeral groups, recommendations must be computed on the basis of those known for the members of the group. A number of different aggregation strategies for the individual preferences have been proposed over the years. The most common examples are: plurality voting (which uses the 'first past the post' principle: repetitively, the item with the most votes is chosen), average (averages individual ratings), additive (sums individual ratings), multiplicative (multiplies individual ratings), Borda count (points are assigned to items based on their ranking, then the item with the most points is chosen), approval voting (counts how many individuals have rating above an approval threshold for a certain item), least misery (takes the minimum of individual ratings), maximum satisfaction (takes the maximum of individual ratings) and most respected person (uses the rating of the most respected individual)[6].

Most of the aggregation strategies just described clearly violate the fairness principles. For instance, *maximum satisfaction*, used in [8, 9, 10, 11, 12, 13], chooses the items for which the greatest value among the preferences of the group members is the highest with the risk of ignoring the satisfaction of most of the users in a group. Another clear example of unfair recommender systems are works such as [14, 15, 16], which assign a different power to group members depending on their expertise.

Fairness in Recommender Systems In single-user Recommender Systems fairness is usually assessed with regard to sensitive attributes which are generally prone to discrimination (e.g., gender, ethnicity or social belonging) by verifying the presence of a discriminated class within the user set [17, 18]. When fairness is evaluated considering Group Recommender Systems, it should be computed within groups. Since the groups we consider in this work are composed of few users, evaluating fairness in the way just described is not a suitable solution. Instead of detecting unfairness towards a protected group of users, we aim to detect and prevent unfairness towards single users within a group whose desires are not taken into consideration when forming a recommendation for the whole group.

Fairness in Group Recommender Systems Some aggregation strategies, that despite not having been developed to explicitly address ethical issues, exist that aggregate individual preferences in a way that resembles fairness. *Least misery*, used in [7, 8, 19, 20, 9, 10, 21, 22, 11, 12, 13, 23], chooses the items for which the lowest value among the preferences of the group members is the greatest. The authors in [24] introduce an aggregation function which tries to maximize the satisfaction of the group components, while, at the same time tries to minimize the *disagreement* among them. The authors in [25] investigate the role played by context in the design of a system that recommends sequences of activities to groups of users as a multi-objective optimization problem, where the satisfaction of the group and the available time interval are two of the functions to be optimized. In particular, their findings suggest that the dynamic evolution of a group can be the key contextual feature that has to be considered to produce fair suggestions. *Average*, used in [8, 26, 27, 19, 20, 9, 10, 11, 14, 12, 13, 23], computes the group preference for an item as the arithmetic mean of the individual scores. Lastly, some recent works try to explicitly target the aim of producing fair group recommendations. In [28] the preferences of individual users are combined with a measure of fairness, to guarantee that all the users are somehow satisfied. In [29, 30] two aggregation strategies are proposed, one is based on the idea of proportionality, while the other one is based on the idea envy-freeness. In [31] the authors formalize the notion of rank-sensitive balance (i.e., an ordered set of recommended items is considered fair to a group if the relevance of the items in the top-N is balanced across the group membes for each prefix of the top-N) and propose a greedy algorithm to achieve it. In our experiments we compare our approach against all the aggregation strategies mentioned in this last subsection.

3 The proposed method

In this section a review of the approach presented in [1, 2], *CtxInfl*, will be given. Then, our contribution to make *CtxInfl* more fair will be presented. The resulting method is named *FARGO*.

3.1 CtxInfl

It is considered a set of items I and a set of users U, from which any group $G \in \wp(U)$ can be extracted. C is the set of possible contexts in the given scenario, where a context c is the conjunction of a set of dimension/value pairs (e.g., for the TV dataset, a context might be $c = \langle time\_slot = primetime \wedge day = weekend \rangle$). It is assumed the availability of a log $\mathcal{L}$ recording the history of the items previously chosen by groups formed in the past, where each element of $\mathcal{L}$ is a 4-ple (t_j, c_j, G_j, i_j) where t_j is the time instant in which the item $i_j \in I$ has been chosen by the group $G_j \in \wp(U)$ in the context $c_j \in C$. A contextual scoring function $score(u, i, c)$, with $u \in U$, $i \in I$, $c \in C$, assigning to each user a score given to the items in the various contexts is computed offline on the basis of the log of the past individual choices and on the basis of the items descriptions

in terms of their attributes using any context-aware recommender system for single users from the literature. $TopK(u, c, t)$ is the function which returns the list of the K items preferred by user u in context c, according to the values of $score(u, i, c)$ for each $i \in I$ available at instant t. Given a target group $G \in \wp(U)$, a context $c \in C$ and a time instant t, the group recommendation is obtained by recommending to the users in G a list (i.e., an ordered set) of K items, considered interesting in context c, from those items in I that are available at time instant t according to the following procedure:

Influence computation The group preference for an item is obtained by aggregating the individual preferences of the group members on the basis of their influence. In each context c, the influence $infl(u, c)$ of a given user u is derived offline by comparing the behavior of u when alone (i.e., u's individual preferences) with u's behaviors in groups (i.e., the interactions contained in the log $\mathcal{L}$). Basically, the influence of u tells how many times the groups containing u have selected one of u's favorite items. Let $TopK(u, c, t)$ be the list of the K items preferred by user u in context c, according to the values of $score(u, i, c)$ for each $i \in I$ available at instant t. The contextual influence is defined as follows:

$$infl(u, c) = \frac{|l_j \in \mathcal{L} : c = c_j \wedge u \in G_j \wedge i_j \in TopK(u, c, t_j)|}{|l_j \in \mathcal{L} : c = c_j \wedge u \in G_j|} \quad (1)$$

The value of $infl(u, c)$ quantifies the ability of user u to direct the group's decision towards u's own tastes while in context c.

Top-K Group Recommendation Computation Top-K recommendations are computed online, when a group of users requires that the system suggests some interesting items to be enjoyed together. The system must compute the group preferences for the items, and then determine the K items with the highest scores. Given a group $G \in \wp(U)$, its preference $score(G, i, c)$ for $i \in I$ in the context $c \in C$ is computed as the average of the preferences of its members weighed on the basis of each member's influence (Eq. 1) in context c:

$$score(G, i, c) = \frac{\sum_{u \in G} infl(u, c) \cdot score(u, i, c)}{\sum_{u \in G} infl(u, c)} \quad (2)$$

Then, the top-K list of items preferred by a certain group G in context c at time instant t is determined by retrieving the K items with the greatest scores among those available at time t.

3.2 FARGO

Being *CtxInfl* based on the concept of influence, it inevitably privileges the preferences of the most influential users. As a consequence, the results of the recommendation process are biased towards the preference of one user or few users of the group which can be considered as the leaders, or, using a more contemporary

word, "influencers". Following the definition of ethics and fairness provided in [32], it is easily understandable that this kind of aggregation strategy doesn't follow any of the fairness principles. Our aim is to add an element of fairness to *CtxInfl* while maintaining its general structure, which already proved to be very efficient and scalable [2]. Among the various phases (i.e., individual preferences computation, influence computation, and Top-K group recommendations computation) of *CtxInfl*, the last one is the most suitable for introducing a fairness element since it is the only one which acts on groups. Following this intuition, we propose to add a fairness factor to the computation of the score for each item (Eq. 2), in order to modify the order of the items in the Top-K list produced so that items which would represent unfair recommendations will not appear on top. This is further motivated by the fact that when people make decisions in groups, they do not always follow the decision of a leader, as assumed by CtxInfl. In some cases people may take decisions trying to satisfy every group member as much as possible. This means that considering only the influence factor may not be a complete strategy even if we put aside our ethical concerns. In order to maintain the computation of Eq. 2 scalable and lightweight we decide to build our fairness element using just the individual contextual scores, which are already needed to compute Eq. 2. We call *consensus* the metric that quantifies how much the individual preferences of the group members agree on the evaluation of an item. The *consensus* of a group G on an item i in a context c is defined as *one* minus the variance of the group members' scores for item i in context c:

$$consensus(G, i, c) = 1 - \frac{\sum_{u \in G} \big(score(u, i, c) - \overline{score}(u, i, c)\big)^2}{|G|} \tag{3}$$

The *consensus* for an item for which users gave a similar evaluation will be close to 1, while it will reach its minimum when very discording scores are considered. According to the formula of the maximum variance:

$$\sigma^2_{max} = \left(\frac{max\big(score(u, i, c)\big) - min\big(score(u, i, c)\big)}{2}\right)^2 = 0.25,$$

$consensus \in [0.75, 1]$, as $score(u, i, c) \in [0, 1]$. After having defined *consensus*, we propose to integrate it in Eq. 2 in the following way:

$$fair\_score(G, i, c) = \frac{\sum_{u \in G} infl(u, c) \cdot score(u, i, c)}{\sum_{u \in G} infl(u, c)} \cdot consensus(G, i, c)^{|G|} \tag{4}$$

We exponentiate *consensus* to the group size (which has the effect of further reducing the overall score) according to the intuition that the magnitude of the problem of unfairness in group recommendations is proportional to the group size. In fact, the bigger is the group, the bigger is the potential harm produced by recommending solely taking into consideration the leader/influencer's will. As an example, given a context $c =$ "*daytime*", an item i available at the time of the recommendation, and a group $G = \{u_1, u_2\}$ composed of two users u_1

and u_2 with contextual influences $infl(u_1, c) = 1$ and $infl(u_2, c) = 0.333$, let's consider the following two cases:

$score(u_1, i, c) = 0.9,\ \ score(u_2, i, c) = 0.2$
$score(G, i, c) = \frac{1 \cdot 0.9 + 0.333 \cdot 0.2}{1 + 0.333} = 0.725$
$consensus(G, i, c) = 1 - \frac{(0.9-0.55)^2 + (0.2-0.55)^2}{2} = 0.8775$
$fair\_score(G, i, c) = 0.725 \cdot 0.8775^2 = 0.558$

$score(u_1, i, c) = 0.7,\ \ score(u_2, i, c) = 0.8$
$score(G, i, c) = \frac{1 \cdot 0.7 + 0.333 \cdot 0.8}{1 + 0.333} = 0.725$
$consensus(G, i, c) = 1 - \frac{(0.7-0.75)^2 + (0.8-0.75)^2}{2} = 0.9975$
$fair\_score(G, i, c) = 0.725 \cdot 0.9975^2 = 0.721$

It can be noted how, even though *CtxInfl* assigns the same group score in both cases, in the case above just the influencer (i.e., u_1) would be truly satisfied with the recommendation. Note also how, in the case below, in which both users like item i, the *consensus* has a minimal impact in the computation of $fair\_score$.

4 Experimental Results

In this section we present the results obtained by applying the proposed approach to two different real-world datasets. To evaluate the recommendation performance we use the *recall*. Let i be an item in the test set, i_t the starting time of availability for i, i_G the group of users that chose the item, i_i the item chosen and i_c the context in which the item was chosen. $TopK(G, c, t)$ indicates the set of top-K items for the group G in context c among those available at time instant t, determined using the recommendation methodology to be evaluated. Recall@K is computed as follows:

$$Recall@K = \frac{|i \in TestSet : i_i \in TopK(i_G, i_c, i_t)|}{|i \in TestSet|}$$

We consider values of K (number of items to be recommended) of 1, 2 and 3.

To evaluate the ethical properties of our method we used the two metrics proposed in [33] for estimating user discrimination, namely, *score disparity* and *recommendation disparity*, which we adapt to our needs.

The first one, called *score disparity*, is computed as the Gini coefficient of user satisfaction (i.e., the relative gain achieved by the user due to the actual recommendation with respect to the optimal recommendation strategy from the user perspective). Firstly, the user satisfaction for a user u is defined as: $\mathcal{A}(u, c, t) = \frac{\sum_{j \in TopK(G,c,t)} score(u,j,c)}{\sum_{j \in TopK(u,c,t)} score(u,j,c)}$, then, *score disparity* is defined as:

$$D_S(G, c, t) = \frac{\sum_{u_1, u_2 \in G} |\mathcal{A}(u_1, c, t) - \mathcal{A}(u_2, c, t)|}{2n \sum_{u \in G} \mathcal{A}(u, c, t)},$$

where n is the number of users.

The second one, called *recommendation disparity*, is computed as the Gini coefficient of user gains (i.e., how many of the the recommended items match the user Top-K items). After computing the user gain with the following formula: $sim(u, c, t) = \frac{|TopK(G,c,t) \cap TopK(u,c,t)|}{K}$, the *recommendation disparity*, is obtained as as:

$$D_R(G, c, t) = \frac{\sum_{u_1,u_2 \in G} |sim(u_1, c, t) - sim(u_2, c, t)|}{2n \sum_{u \in G} sim(u, c, t)}.$$

The choice of these specific metrics for quantifying fairness relies on the fact that they have been proposed to specifically consider the disparate impacts of recommendations on different users.

We compare our approach to the following methods: average (**AVG**) [8, 26, 27, 19, 20, 9, 10, 11, 14, 12, 13, 23], **Fair Lin** [28], **Fair Prop** [29, 30], **Envy Free** [29, 30], minimum disagreement (**Dis**) [24], least misery (**LM**) [7, 8, 19, 20, 9, 10, 21, 22, 11, 12, 13, 23] and **GFAR** [31]. The choice of these specific methods is motivated by the fact that they are either very well-known and broadly adopted baselines (e.g., AVG, LM) or represent recent state-of-the-art approaches for fair group recommendations (e.g., Fair Lin, Fair Prop, Envy Free, GFAR).

4.1 TV Dataset

This dataset contains TV viewing information related to 7,921 users and 119 channels, broadcasted both over the air and by satellite. The dataset is composed of an electronic program guide (EPG) containing the description of 21,194 distinct programs, and a log containing both individual and group viewings performed by the users. The log spans from December 2012 to March 2013 and contains 4,968,231 entries, among which we retained just the syntonizations longer than three minutes. 3,519,167 viewings are performed by individual users, which are used to compute the individual preferences of the group members. The remaining 1,449,064 viewings have been done by more than one person. The two context dimensions considered in the experiments are *day of the week* (weekday vs. weekend) and the *time slot*. The available values for the time slot are: graveyard slot (from 02:00 to 07:00), early morning (from 07:00 to 09:00), morning (from 09:00 to 12:00), daytime (from 12:00 to 15:00), early fringe (from 15:00 to 17:00), prime access (from 18:00 to 20:30), primetime (from 20:30 to 22:30), and late fringe (from 22:30 to 02:00). Group viewings are split into a training set (1,210,316 entries), and a test set (238,748 entries) with a 80%-20% ratio. This way of splitting the dataset is selected in order to maintain the usual Recommender System literature splitting percentage: 80% training set, 20% test set. Results are reported in Table 1 considering $K = 1$, $K = 2$ and $K = 3$.

The superiority of our method recall-wise is very pronounced. For what regards the ethical guarantees, *FARGO*, delivers a very good *score disparity*, while, for what regards the *recommendation disparity*, it seems to perform generally worse than the other methods (except for $k = 1$, for which its performance is on par with the other methods). Please note that for GFAR it is not possible to

	K=1			K=2			K=3		
	Recall	**D_S**	**D_R**	**Recall**	**D_S**	**D_R**	**Recall**	**D_S**	**D_R**
FARGO	**37.94%**	7.61%	17.85%	**54.08%**	**1.85%**	12.69%	**64.20%**	**0.89%**	10.08%
AVG	33.914%	7.07%	18.15%	51.56%	2.93%	8.78%	62.91%	1.36%	7.53%
Fair Lin	33.22%	8.83%	18.25%	50.80%	3.59%	7.46%	61.21%	1.61%	**7.01%**
Fair Prop	32.99%	8.83%	13.45%	50.55%	4.25%	8.90%	62.03%	1.79%	7.70%
Envy Free	29.33%	10.43%	13.81%	47.37%	4.23%	10.87%	58.67%	1.89%	8.72%
Dis	33.57%	6.67%	17.45%	51.95%	2.76%	8.97%	63.26%	1.30%	7.61%
LM	30.35%	**5.69%**	**12.42%**	47.10%	2.58%	10.11%	58.27%	1.25%	8.18%
GFAR	30.47%	-	18.28%	44.48%	-	**5.59%**	55.19%	-	7.41%

Table 1: Comparison with other fair methods on TV dataset

	K=1			K=2			K=3		
	Recall	**D_S**	**D_R**	**Recall**	**D_S**	**D_R**	**Recall**	**D_S**	**D_R**
FARGO	**25.00%**	2.19%	**1.62%**	**40.28%**	0.87%	2.03%	**49.31%**	0.53%	2.40%
AVG	12.50%	0.81%	3.24%	25.00%	0.39%	2.91%	34.72%	**0.25%**	2.71%
Fair Lin	11.11%	2.19%	4.17%	23.61%	0.81%	2.14%	31.94%	0.48%	**1.81%**
Fair Prop	13.19%	**0.66%**	2.55%	20.83%	0.38%	3.24%	29.86%	0.37%	3.00%
Envy Free	12.50%	0.81%	3.24%	25.00%	0.39%	2.95%	34.72%	**0.25%**	2.71%
Dis	22.92%	0.74%	3.41%	34.72%	0.43%	2.83%	41.67%	0.32%	2.49%
LM	13.89%	1.14%	3.76%	25.00%	**0.35%**	3.13%	34.72%	0.28%	1.99%
GFAR	6.06%	-	8.73%	24.24%	-	**1.88%**	33.33%	-	6.24%

Table 2: Comparison with other fair methods on Music dataset

compute the score disparity as the method does not involve the computation of group scores for the items.

4.2 Music Dataset

This dataset has been created by asking participants to fill in two different forms: an *individual form* collecting demographic data (i.e., age and gender) and contextual individual preferences about music artists, and a *group form* to be filled in groups asking for a collective choice of a music artist that was available at the time of the choice in a particular context. The following two listening contexts have been selected considering that both are common situations users can relate to both when alone and when with other people, and that users' preferences would likely be different in each of them: *during a car trip* and *at dinner as background music*. We defined a list of 30 music artists well-known in Italy covering most of the music genres available. The metric used to evaluate preferences is a number between 0 and 4 reflecting the following list:

0. user would not listen to it or user does not know it
1. user would listen to it very seldom
2. user would listen to it sometimes
3. user would listen to it often

4. user would always listen to it

The dataset obtained contains data gathered from 280 users. The user set is composed by 57 females and 223 males, the age of the users is between 18 and 60. Since the forms have been proposed mostly to university students the users' average age is in the interval 18-30. For each user, preferences regarding both the car trip and dinner contexts are gathered. From the group forms, 498 context-aware collective preferences have been gathered. Of this, 272 groups were composed of 2 users, 158 of 3 users, 32 of 4 users and 36 of 5 users. As for the previous dataset, we used a 80%-20% split for training and test sets. This dataset has been collected specifically for this work and is made publicly available as one of the contributions of this article. The dataset can be downloaded at `https://github.com/azzada/FARGO`.

Results are reported in Table 2. Also in this case *FARGO* delivers the best recall. Contrarily to the previous dataset, in this case our method achieves a very good *recommendation disparity*. For what regards the *score disparity*, all methods provide very low (i.e., good) values.

5 Conclusions

In this paper we have introduced FARGO, a new method for providing fair, context-aware recommendations to ephemeral groups, which is also able to recommend items that are new in the system. If we consider both recall and fairness, it is not possible to identify a best overall method across all datasets and values of K. Even if we ignored recall, a clear winner fairness-wise is not evident (all of tested methods, except for *Dis*, perform best fairness-wise for at least a value of K in at least one of the two datasets). We argue that the relationship between fairness and recommendation accuracy should be seen as a tradeoff. On both datasets of our experiments, FARGO provides the best solution to such tradeoff by achieving the best recall across all values of K while delivering similar ethical guarantees to the other fair methods tested. Contrarily to what one might think, *LM* is not the best method fairness-wise, this implies that the problem of maximizing both recall and fairness is not a simple one. This is a complex problem that deserves further investigations, as recall and fairness seem not to be inversely correlated in a trivial manner. Although there may be some combinations of individual preference scores that would lead FARGO to violate the principle of social choice (imagine that there are three users, A, B and C and two items i and j; if the individual scores for i are 0.4, 0.4 and 0.4, while those for j are 0.4, 0.4 and 0.6, then the proposed method, depending also on the influences, may suggest item i, which has a higher consensus, despite the fact that clearly item j would make all users equally or better satisfied), it turns out that this is actually not a problem in practice as by filtering out at runtime dominated items (and hence respecting the social choice principle) both recall and fairness metrics worsen significantly. This may suggest that in reality groups tend to choose items that make everyone equally happy.

Future works include better investigating the reasons why the two fairness measures are somewhat unstable between the two datasets (e.g., investigating why *FARGO* delivers a better score disparity for the TV dataset, while, for the Music dataset, it delivers a better recommendation disparity), as well as finding alternatives to the consensus as a fairness element to be integrated in *CtxInfl*.

References

[1] Elisa Quintarelli, Emanuele Rabosio, and Letizia Tanca. "Recommending new items to ephemeral groups using contextual user influence". In: *Proc. RecSys*. 2016, pp. 285–292.

[2] Elisa Quintarelli, Emanuele Rabosio, and Letizia Tanca. "Efficiently using contextual influence to recommend new items to ephemeral groups". In: *Inf. Syst.* 84 (2019), pp. 197–213.

[3] Francesco Ricci et al. *Recommender Systems Handbook*. Springer, 2011.

[4] Gediminas Adomavicius and Alexander Tuzhilin. "Context-Aware Recommender Systems". In: *Recommender Systems Handbook*. Springer, 2011, pp. 217–253.

[5] Katrien Verbert et al. "Context-Aware Recommender Systems for Learning: A Survey and Future Challenges". In: *IEEE Transactions on Learning Technologies* 5.4 (2012), pp. 318–335.

[6] Judith Masthoff. "Group Recommender Systems: Combining Individual Models". In: *Recommender Systems Handbook*. Ed. by Francesco Ricci et al. Springer, 2011, pp. 677–702.

[7] Mark O'Connor et al. "PolyLens: A Recommender System for Groups of Users". In: *Proc. ECSCW*. 2001, pp. 199–218.

[8] Judith Masthoff. "Group modeling: Selecting a sequence of television items to suit a group of viewers". In: *Personalized Digital Television*. Springer, 2004, pp. 93–141.

[9] Steven Bourke, Kevin McCarthy, and Barry Smyth. "Using social ties in group recommendation". In: *Proc. 22nd Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*. 2011.

[10] Eirini Ntoutsi et al. "Fast group recommendations by applying user clustering". In: *Proc. ER*. 2012, pp. 126–140.

[11] Allison JB Chaney et al. "A large-scale exploration of group viewing patterns". In: *Proc. TVX*. 2014, pp. 31–38.

[12] Toon De Pessemier, Simon Dooms, and Luc Martens. "Comparison of group recommendation algorithms". In: *Multimedia Tools Appl.* 72.3 (2014), pp. 2497–2541.

[13] Noo-ri Kim and Jee-Hyong Lee. "Group recommendation system: Focusing on home group user in TV domain". In: *Proc. SCIS*. 2014, pp. 985–988.

[14] Irfan Ali and Sang-Wook Kim. "Group recommendations: approaches and evaluation". In: *Proc. IMCOM*. 2015, pp. 1–6.

[15] Mike Gartrell et al. "Enhancing group recommendation by incorporating social relationship interactions". In: *Proc. GROUP*. 2010, pp. 97–106.

[16] Shlomo Berkovsky and Jill Freyne. "Group-based recipe recommendations: analysis of data aggregation strategies". In: *Proc. RecSys*. 2010, pp. 111–118.
[17] Sirui Yao and Bert Huang. "New Fairness Metrics for Recommendation that Embrace Differences". In: *CoRR* abs/1706.09838 (2017).
[18] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. "Tutorial on Fairness of Machine Learning in Recommender Systems". In: *Proc. SIGIR*. 2021, pp. 2654–2657.
[19] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. "Group recommendations with rank aggregation and collaborative filtering". In: *Proc. RecSys*. 2010, pp. 119–126.
[20] Christophe Senot et al. "Analysis of strategies for building group profiles". In: *Proc. UMAP*. 2010, pp. 40–51.
[21] Oskar Van Deventer et al. "Group recommendation in a hybrid broadcast broadband television context". In: *Group Recommender Systems: Concepts, Technology, Evaluation* 997 (2013), pp. 12–18.
[22] Jagadeesh Gorla et al. "Probabilistic group recommendation via information matching". In: *Proc. WWW*. 2013, pp. 495–504.
[23] Sarik Ghazarian and Mohammad Ali Nematbakhsh. "Enhancing memory-based collaborative filtering for group recommender systems". In: *Expert Syst. Appl.* 42.7 (2015), pp. 3801–3812.
[24] Sihem Amer-Yahia et al. "Group recommendation: Semantics and efficiency". In: *Proc. VLDB*. 2009, pp. 754–765.
[25] Sara Migliorini et al. "What is the Role of Context in Fair Group Recommendations?" In: *Proc. PIE@CAiSE*. 2019.
[26] Zhiwen Yu et al. "TV program recommendation for multiple viewers based on user profile merging". In: *User Model. User-Adapt. Int.* 16.1 (2006), pp. 63–82.
[27] Choonsung Shin and Woontack Woo. "Socially aware TV program recommender for multiple viewers". In: *IEEE Trans. on Consumer Electronics* 55.2 (2009), pp. 927–932.
[28] Lin Xiao et al. "Fairness-Aware Group Recommendation with Pareto-Efficiency". In: *Proc. RecSys*. 2017, pp. 107–115.
[29] Shuyao Qi et al. "Recommending packages to groups". In: *Proc. ICDM*. 2016, pp. 449–458.
[30] Dimitris Serbos et al. "Fairness in Package-to-Group Recommendations". In: *Proc. WWW*. 2017, pp. 371–379.
[31] Mesut Kaya, Derek Bridge, and Nava Tintarev. "Ensuring fairness in group recommendations by rank-sensitive balancing of relevance". In: *Proc. RecSys*. 2020, pp. 101–110.
[32] Cynthia Dwork et al. "Fairness through awareness". In: *Proc. ITCS*. 2012, pp. 214–226.
[33] Jurek Leonhardt, Avishek Anand, and Megha Khosla. "User Fairness in Recommender Systems". In: *Proc. WWW*. 2018, pp. 101–102.