Accounting for Bossy Users in Context-Aware Group Recommendations

(Discussion Paper)

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*

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 are not appropriate and, as a consequence, over the years a number of group recommender systems have been developed. Recommending items to be enjoyed together by a group of people poses many ethical challenges: in fact, a system whose unique objective is to achieve the best recommendation accuracy 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 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 to 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, while the second one gathers the musical preferences (both individual and in groups) of 280 real users w.r.t. two contexts of listening. 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

1. Introduction

Recommender Systems are software tools and techniques that provide suggestions for items to be of use to a user [1]. 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 any history of the group's preferences available, 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: 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, e.g. 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 [2, 3] to account also for fairness. Experiments on two real-world datasets show that our approach outperforms seven other fair group recommender systems by achieving a consistently better recall while providing similar ethical guarantees.

2. Related Work

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, the context of use might be fundamental in guiding the current preference of a user [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 that produce a common recommendation for a group of users [6]. Group recommendations works usually address two kinds of groups: *persistent* and *ephemeral* [7]. Persistent groups contain people that have a previous significant history of activities together, while ephemeral groups are formed by people who 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 [6], however most of these aggregation strategies clearly violate the fairness principles. For instance, *maximum satisfaction*, used in [8, 9, 10, 11, 12], chooses the item for which the individual preference score is the highest, effectively ignoring the satisfaction of most of the users in the group. Other clear examples of unfair aggregation strategies are works such as [13, 14, 15], which assign a different power to group members based 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 [16, 17]. 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 considered when forming a recommendation for the whole group.

Fairness in Group Recommender Systems

Some aggregation strategies exist that, despite not having been developed to explicitly address ethical issues, aggregate individual preferences in a way that resembles fairness. *Least misery*, used in [7, 8, 18, 19, 9, 20, 10, 11, 12, 21], chooses the items for which the lowest value among the preferences of the group members is the greatest one. The authors in [22] introduce an aggregation function which maximizes the satisfaction of group components, while, at the same time, minimizes the *disagreement* among them. *Average*, used in [8, 23, 18, 19, 9, 10, 13, 11, 12, 21], 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 [24] the preferences of individual users are combined with a measure of fairness, to guarantee that all the users are somehow satisfied. In [25, 26] 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 [27] a greedy algorithm to achieve rank-sensitive balance is presented.

3. The proposed method

In this section we review a previous approach of ours, introduced in [2, 3], *CtxInfl*. Then, our contribution to make *CtxInfl* more fair will be presented. The resulting method is named *FARGO*.

3.1. CtxInfl

We considere 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 $c = \langle time\_slot = primetime \wedge day = weekend \rangle$. We assume 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), t_j being 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 the score given to the items in the various contexts, is computed offline on the basis of the log of the past individual choices and of the item 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 that 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:

3.1.1. 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 us 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:

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

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.

3.1.2. 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:

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

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 highest 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 who can be considered as the leaders, or, using a more contemporary word, "influencers". 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 [3]. Among the various phases of *CtxInfl* on which we could act (i.e., individual preferences computation, influence computation, and Top-K group recommendations computation), the last is the most suitable one, as it is the only one acting 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 in such a way that items representing unfair recommendations will not appear on top. This is further motivated by the fact that, when people make decisions in groups, not necessarily they 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 influence may not be a complete strategy even if we put aside our ethical concerns. In order not to increase the complexity of the computation of Eq. 2, we build our fairness element using just the individual contextual scores, which are already used to compute Eq. 2. We call *consensus* the metric that quantifies how much the individual preferences of group members agree on the evaluation of an item. The *consensus* of a group G on an item i in a context c is therefore defined as:

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

where $\overline{score}(u, i, c)$ is the average score for item i among G's members in context c. 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 discordant scores are considered. According to the formula of the maximum variance, $consensus \in [0.75, 1]$. 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 (with the effect of further reducing the overall score) according to the intuition that the magnitude of unfairness in group recommendations is proportional to the group size. In fact, the bigger the group, the bigger the potential harm produced by taking into consideration solely the leader/influencer's will is.

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 recall, considering for K (number of items to be recommended) the values 1, 2 and 3. To evaluate the ethical properties of our method we used the two metrics proposed in [28] for estimating user discrimination, called *score disparity* and *recommendation disparity*, adapted to our needs. *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. *Recommendation disparity* is computed as the Gini coefficient of user gains, i.e., how many of the recommended items match the user Top-K items.

We compare our approach to the following methods: average (**AVG**) [8, 23, 18, 19, 9, 10, 13, 11, 12, 21], **Fair Lin** [24], **Fair Prop** [25, 26], **Envy Free** [25, 26], minimum disagreement (**Dis**) [22], least misery (**LM**) [7, 8, 18, 19, 9, 20, 10, 11, 12, 21] and **GFAR** [27].

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

	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

February 2013 and contains 4,968,231 entries, among which we retained just the syntonizations longer than three minutes. 3,519,167 viewings were performed by individual users, and 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 are *day of the week* (weekday vs. weekend) and the *time slot*. The available values for the time slot are: graveyard slot, early morning, morning, daytime, early fringe, prime access, primetime, and late fringe. Group viewings are split into a training set (1,210,316 entries), and a test set (238,748 entries) with a 80%-20% ratio. Results are reported in Table 1. Note that 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). Note that for GFAR it is not possible to compute the score disparity as it does not involve the computation of group scores for the items.

4.2. Music Dataset

This dataset[1] has been created within the scope of a user study 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*. The dataset obtained contains data gathered from 280 users. 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. 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.

---

[1]The dataset can be downloaded at https://github.com/azzada/FARGO.

	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

5. Conclusions

In this paper we have introduced FARGO, a new method for providing fair, context-aware recommendations to ephemeral groups able also to recommend items new in the system. Considering 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 methods tested, except for *Dis*, perform best fairness-wise for at least a value of K in at least one dataset). 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, and 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.

References

[1] F. Ricci, L. Rolach, B. Shapira, P. B. Kantor, Recommender Systems Handbook, Springer, 2011.

[2] E. Quintarelli, E. Rabosio, L. Tanca, Recommending new items to ephemeral groups using contextual user influence, in: Proc. RecSys, 2016, pp. 285–292.

[3] E. Quintarelli, E. Rabosio, L. Tanca, Efficiently using contextual influence to recommend new items to ephemeral groups, Inf. Syst. 84 (2019) 197–213.

[4] G. Adomavicius, A. Tuzhilin, Context-Aware Recommender Systems, Springer, 2011, pp. 217–253.

[5] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, E. Duval, Context-aware recommender systems for learning: A survey and future challenges, IEEE Transactions on Learning Technologies 5 (2012) 318–335.

[6] J. Masthoff, Group Recommender Systems: Combining Individual Models, Springer, 2011, pp. 677–702.

[7] M. O'Connor, D. Cosley, J. A. Konstan, J. Riedl, Polylens: A recommender system for groups of users, in: Proc. ECSCW, 2001, pp. 199–218.

[8] J. 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] E. Ntoutsi, K. Stefanidis, K. Nørvåg, H.-P. Kriegel, Fast group recommendations by applying user clustering, in: Proc. ER, 2012, pp. 126–140.

[10] A. J. Chaney, M. Gartrell, J. M. Hofman, J. Guiver, N. Koenigstein, P. Kohli, U. Paquet, A large-scale exploration of group viewing patterns, in: Proc. TVX, 2014, pp. 31–38.

[11] T. De Pessemier, S. Dooms, L. Martens, Comparison of group recommendation algorithms, Multimedia Tools Appl. 72 (2014) 2497–2541.

[12] N.-r. Kim, J.-H. Lee, Group recommendation system: Focusing on home group user in tv domain, in: Proc. SCIS, 2014, pp. 985–988.

[13] I. Ali, S.-W. Kim, Group recommendations: approaches and evaluation, in: Proc. IMCOM, 2015, pp. 1–6.

[14] M. Gartrell, X. Xing, Q. Lv, A. Beach, R. Han, S. Mishra, K. Seada, Enhancing group recommendation by incorporating social relationship interactions, in: Proc. GROUP, 2010, pp. 97–106.

[15] S. Berkovsky, J. Freyne, Group-based recipe recommendations: analysis of data aggregation strategies, in: Proc. RecSys, 2010, pp. 111–118.

[16] S. Yao, B. Huang, New fairness metrics for recommendation that embrace differences, CoRR abs/1706.09838 (2017).

[17] Y. Li, Y. Ge, Y. Zhang, Tutorial on fairness of machine learning in recommender systems, in: Proc. SIGIR, 2021, pp. 2654–2657.

[18] L. Baltrunas, T. Makcinskas, F. Ricci, Group recommendations with rank aggregation and collaborative filtering, in: Proc. RecSys, 2010, pp. 119–126.

[19] C. Senot, D. Kostadinov, M. Bouzid, J. Picault, A. Aghasaryan, C. Bernier, Analysis of strategies for building group profiles, in: Proc. UMAP, 2010, pp. 40–51.

[20] J. Gorla, N. Lathia, S. Robertson, J. Wang, Probabilistic group recommendation via information matching, in: Proc. WWW, 2013, pp. 495–504.

[21] S. Ghazarian, M. A. Nematbakhsh, Enhancing memory-based collaborative filtering for group recommender systems, Expert Syst. Appl. 42 (2015) 3801–3812.

[22] S. Amer-Yahia, S. B. Roy, A. Chawlat, G. Das, C. Yu, Group recommendation: Semantics and efficiency, in: Proc. VLDB, 2009, pp. 754–765.

[23] Z. Yu, X. Zhou, Y. Hao, J. Gu, Tv program recommendation for multiple viewers based on user profile merging, User Model. User-Adapt. Int. 16 (2006) 63–82.

[24] L. Xiao, Z. Min, Z. Yongfeng, G. Zhaoquan, L. Yiqun, M. Shaoping, Fairness-aware group recommendation with pareto-efficiency, in: Proc. RecSys, 2017, pp. 107–115.

[25] S. Qi, N. Mamoulis, E. Pitoura, P. Tsaparas, Recommending packages to groups, in: Proc. ICDM, 2016, pp. 449–458.

[26] D. Serbos, S. Qi, N. Mamoulis, E. Pitoura, P. Tsaparas, Fairness in package-to-group recommendations, in: Proc. WWW, 2017, pp. 371–379.

[27] M. Kaya, D. Bridge, N. Tintarev, Ensuring fairness in group recommendations by rank-sensitive balancing of relevance, in: Proc. RecSys, 2020, pp. 101–110.

[28] J. Leonhardt, A. Anand, M. Khosla, User fairness in recommender systems, in: Proc. WWW, 2018, pp. 101–102.