Top-N Group Recommendations with Fairness

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

ABSTRACT

In many settings it is required that items are recommended to a group of users instead of a single user. Conventionally, the objective is to maximize the overall satisfaction among group members. Recently, however, attention has shifted to ensuring that recommendations are fair in that they should minimize the feeling of dissatisfaction among members. In this work, we explore a simple but intuitive notion of fairness: the minimum utility a group member receives. We propose a technique that seeks to rank the Pareto, or unanimously, optimal items by considering all admissible ways in which a group might reach a decision. As our detailed experimental study shows, this results in top-N recommendations that not only achieve a high minimum utility compared to other fairness-aware techniques, but also a high average utility across all group members beating standard aggregation strategies.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Group Recommender Systems, Aggregation Strategies, Fairness, Pareto Efficiency

ACM Reference Format:
Dimitris Sacharidis. 2019. Top-N Group Recommendations with Fairness. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3297280.3297442

1 INTRODUCTION

The typical role of a recommender system is to suggest items to individual users for consumption based on their preferences. In many settings however, a recommendation to a group of people rather than to a single user is required, e.g., to friends planning their summer vacation destination [1], or to a family deciding on a TV program to watch [32]. The added difficulty here is that recommendations should satisfy all group members, which can have diverse or even conflicting preferences. The absence of (or difficulty of obtaining) any information about the group's decision process and dynamics, which is often the case with *ad hoc*, or ephemeral, groups, further complicates the job of the recommender. Therefore, one of the primary challenges for group recommenders is to decide what the group preferences are, or more precisely how to derive them from individual preferences.

We discern two basic paradigms for providing group recommendations. The first is to *aggregate profiles* of individual members, e.g., their rating history in a pure collaborative filtering setting, so as to construct a group profile. In this way, the group can be treated as a virtual user, and standard techniques to provide recommendations can be employed. The second paradigm is to *aggregate recommendations* compiled for each member separately. Inspired by social choice theory, numerous aggregation strategies for profiles and recommendations have been used [14]. For example, the average strategy assigns to an item a group rating calculated as the average of (predicted) member ratings, whereas the least-misery strategy assigns the lowest member rating. In this work, we focus on this latter paradigm.

In the vast majority of work on group recommenders, the canonical objective is to maximize the group's overall satisfaction with the recommended list. However, more recently, there has been great interest in making recommendations that are *fair* to each group member [11, 21, 29]. In this context, fairness attempts to minimize the feeling of dissatisfaction within group members. In this work, we propose a simple but intuitive definition of fairness. Suppose that we have a measure of quantifying the satisfaction, or *utility*, that a group member receives from a list of recommendations. Then, for a specific list of recommendations, we can define the *group utility* to be the average member utility (e.g., the social welfare according to [11]), and *fairness* to be the minimum member utility. Intuitively, a list that minimizes the dissatisfaction of any group member should be regarded as the most fair. In this sense, fairness enforces the least misery principle among member utilities [11].

The aforementioned notions of group utility and fairness critically depend on how a member's individual utility is quantified. Intuitively, we want to assign a high utility whenever a list of group recommendations satisfies, i.e., closely matches, the member's individual preferences. Assume that the system can extract the top-N items to individually recommend to this member; e.g., using any collaborative filtering technique for top-N recommendations. This top-N list is considered the *ground truth* for this member, achieving the highest possible utility (to the best of the system's knowledge of the user's preferences). Then, the utility of any other list should be computed relative to this ground truth. Therefore, for a particular group member we define her *utility* of a list of group recommendations as the similarity between the list and the member's personal ground truth. We note that as there exist several natural measures that quantify similarity between two lists, including symmetrical

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04…$15.00
https://doi.org/10.1145/3297280.3297442

ones, e.g., Spearman's footrule or rho, Kendall's tau, and unsymmetrical ones, e.g., precision, recall, average precision, normalized discounted cumulative gain (ndcg), any among them could be used to define a member's utility.

We remark that our definition of utility differs from that in [11], where the member's utility to a list of recommendations is equal to the sum of relevances (e.g., explicit ratings) of each item in the list. This implies that each item has the same contribution to the member's utility irrespectively of its position in the list. As a result, a list and its reverse are considered to have the same utility. This makes sense when recommending packages to groups (e.g., [21, 29]), but is counterintuitive for top-N recommendations. More interestingly, however, our definition of utility essentially generalizes an implicit definition made in [2]. There, the quality of a group recommender is quantified in terms of the average ndcg that the list achieves for each member. In other words, each member has a ground truth list compiled from her ratings, and the utility of a list to a member is given by the ndcg with respect to her ground truth.

Our approach for making fair top-N group recommendations is based on the notion of Pareto optimality. An item is *Pareto optimal* for a group if there exists no other item that ranks higher according to all group members. In other words, there exists no item that is considered by the group *unanimously* better than a Pareto optimal item. The set of Pareto optimal items is inherently *fair* in the sense that it includes the top choice for each group member. In addition, it includes other items, besides the top-1's, that represent different trade-offs among members; e.g., an item that is the second best for all members might be a better option than an item ranking first for just one member. The set of Pareto optimal items is also *complete* in the sense that it contains only those items that can be the top choice of any rational (monotonic with respect to member's preferences) group decision.

To compile a list of top-N recommendations, we consider among items in the set of N-level Pareto optimal items, a notion that extends the aforementioned properties for the top-1 choice to the case of top-N choices. This ensures that no member is treated unfairly, as each has an equal chance to contribute to the group recommendations. However, we also need to ensure that we select N items so that they represent good choices for each member. Thus, we propose a simple method that assigns to each item in the Pareto optimal set a score proportional to its probability of being within the top-N choices among all possible rational group decisions. Intuitively, we seek items that represent good compromises among the group members, and the higher this probability is for an item, the better its compromise is expected to be. The list of top-N group recommendations consists of the N items that have the highest such score.

We propose two variants on this basic idea. The first makes N recommendations among the set of N-level Pareto optimal items. The second, first selects a level $k < N$ such that the number of k-level Pareto optimal items is at least N, and then chooses the best N among them.

We perform a detailed evaluation study using synthetically generated groups of users based on the MovieLens dataset. We create groups of 2 up to 8 random or similar users, and request 5 up to 100 group recommendations. We compare our methods against standard recommendation aggregation methods, as well as fairness-aware techniques. Our results demonstrate that our Pareto-based recommenders produce not only recommendations that significantly more fair, but that also have higher group utility, in almost all settings tested. The gains in fairness of our methods are more pronounced in settings that matter the most in practice, e.g., in the first few ranks and for medium-sized groups. Other notable conclusions from our evaluation include the following. As the size of the group increases, it becomes harder to make fair recommendations to them, especially for groups of random users. On the other hand, as the number of requested recommendations increases, it becomes easier to produce fair recommendations.

The remainder of this paper is organized as follows. Section 2 presents related work. Then Section 3 defines the problem and introduces our approach, and Section 4 presents the results of our evaluation study. Section 5 concludes with general observations.

2 RELATED WORK

Literature on group recommenders is rich; we refer the reader to [10, 14] for a systematic treatment of this research area.

One important distinction is whether groups are *persistent* or *ephemeral*. In the former case, e.g., a household, there is typically enough historical information about group-item interactions to treat the group as a virtual user for whom recommendations are made. In the latter case, groups are formed ad-hoc, e.g., a bunch of friends arranging a dinner, and thus there no historical data about how group-item interactions. There is also the case in-between of a few observed group-item interactions, where techniques [5, 23, 26] try to make use of both member-item and group-item interactions. In this work, we deal with ephemeral groups.

There exist two basic paradigms for providing recommendations to ephemeral groups. In *profile aggregation*, also referred to as aggregated model [3] or group model [13], a group profile is created by aggregating the profiles of group members, see e.g., MusicFX [16], Yu's TV recommender [32], and the content-based TV recommender in [28]. In this way, the group can again be treated as a virtual user, and standard recommendation techniques apply. The second paradigm is to *aggregate recommendations* compiled for each member separately. Inspired by social choice theory, numerous aggregation strategies for profiles and recommendations have been used [14].

We further distinguish two classes of strategies for aggregating recommendations. In the first, termed *rating aggregation*, an item is explicitly assigned a group rating determined by an aggregation over the predicted member ratings. The aggregation strategies are mostly inspired by social choice theory (see [13] for an overview), and include taking the average, the minimum a.k.a. least misery principle of not strongly displeasing any member, the maximum for ensuring the greatest pleasure among members, and the product. The vast majority of past work falls in this category; for reference we mention the following systems: POLYLENS [17] that aggregates recommendations assuming least misery, and INTRIGUE [1] that is an interesting hybrid that first identifies sub-groups among groups (e.g., children, or disabled persons) and creates an aggregate profile for each, and then aggregates recommendations using a weighted averaging scheme.

In the second class, termed *rank aggregation*, the position of an item in the outputted group recommendation list is explicitly determined. Here the inspiration comes from rank aggregation techniques for top-k lists [6, 7]. The work in [2] introduced this approach for group recommenders, where Borda count and Spearman footrule aggregation were used.

It has been observed, e.g., in [15], that group dynamics play a key role in group decisions. Therefore, several group recommenders take also into account the group composition. For example, [8, 24, 25] study the social connections among members, while [22], [3], and [12] consider the personality traits, roles, and authority on topics, respectively, of group members. This line of work however does not apply in our setting, where we make minimal assumptions about what information is available to the recommender. Specifically, we only assume member-item interactions are known.

While a significant focus on group recommenders research is on the group's satisfaction, there is some recent work that seeks to ensure fairness in group recommendations. A simple idea is to include in the aggregation function a penalty term that measures the amount of variation in the predicted ratings among the group members [10]. [30] uses the median of predicted member ratings, so as "to cause the lowest overall change of the individual user preference". Fairness for groups can also be considered as a multi-stakeholder recommendation problem, where each member and the whole group can be viewed as different stakeholders; fairness in such settings is discussed in [4].

For the problem of recommending packages, i.e., a set of items instead of a single item, to groups, the work in [20] considers a fairness measure when compiling packages. Specifically, a package is fair to a group member if it contains at least one item that the member ranks with her top-N items. The goal of the system is to identify the best package, where each package is assessed by a score that is the product of the group utility of its items and the ratio of users that find it fair. We refer to the greedy algorithm in [20] as GRF. For the same problem of package to group recommendations, [29] extends the previous fairness measure to *m-proportionality*, which is the ratio of members that find at least m items in the package to be within their top-N choices. Moreover, the concept of *m-envy-freeness* is introduced, which is the ratio of members that find at least m items in the package so that their predicted rating for each item is among the best $\Delta\%$ ratings the item is predicted to receive by the other members. [29] introduces greedy algorithms, which we denote as SPG and EFG, to maximize m-proportionality and m-envy-freeness, respectively.

For the problem of top-N recommendations to groups, [11] defines the utility of a member to a list as the (normalized) sum of predicted relevance values (e.g., ratings) of all items in the list. Then, the group utility (termed social welfare) is the average member utility. Fairness is considered in four flavors that operate over member utilities. *Least misery fairness* is the minimum utility of a member; *variance fairness* is the negative variance of the member utilities; *min-max ratio fairness* is the ratio of the minimum to maximum member utility; *Jain's fairness* is the ratio of the average squared member utility over the square of the average member utility. The objective is to create top-N recommendations that have both high group utility and fairness. The best performing algorithm is a greedy method, which we denote as GVAR, that optimizes for variance fairness.

A significant line of work concerns the evaluation of group recommenders. The seminal work of [15] studies what factors influence group satisfaction and how it differs from individual satisfaction. A comparative evaluation of profile aggregation strategies can be found at [27], while studies of profile aggregation and rating aggregation methods are presented in [3, 18]. We also note that [2] compares rank aggregation techniques with rating aggregation but finds no clear winner.

3 APPROACH

Problem Definition. For the remaining of this section, we consider a particular group g consisting of m users. The only requirement from the system is that it can generate top-N recommendations for each group member. We refer to the top-N list for a member as her *ground truth.* The *member utility* of a list of group recommendations is equal to a similarity measure between the list and the member's ground truth. The *group utility* of a list is the average member utility. The *fairness* of a list is the lowest member utility. The problem is to compile a list of top-N recommendations to the group so as to maximize fairness.

Preliminaries. Any item i can be described as a vector in $\mathbb{R}^m$ space, where each dimension corresponds to a group member, say u, and its coordinate equals the rank $r_u(i)$ of the item for this member. Consider the example shown in Figure 1 about a group of two users, u_1 and u_2. The top-6 items for user u_1 are $i_2, i_3, i_5, i_1, i_6, i_4$, while for u_2 are $i_1, i_4, i_2, i_3, i_6, i_5$. Item i_1 ranks fourth for one member and first for the other, and is thus represented by the point $(4, 1)$ on the plane. Observe that it is often impossible to compare two items, e.g., i_1 and i_2 as they represent different trade-offs among the group members. In some case, however, one item, e.g., i_1, can be clearly better than another, e.g., i_4.

We say that item i *dominates* another i' according to group g, if for each member item i ranks at least as good as i', and there exists at least one member for whom i ranks better, i.e., $\forall u \in g : r_u(i) \leq r_u(i')$, and $\exists u' \in g : r_{u'}(i) < r_{u'}(i')$; we assume that ties in ranks are possible. Intuitively, dominance means that the group *unanimously* agrees that i is a better item than i', as is the case with i_1 dominating i_4 in the example.

An aggregation strategy is Pareto-efficient, or simply *Pareto*, if whenever every member ranks one item higher than another, then so does the strategy [31]. In other words, a Pareto aggregation strategy respects the dominance relation among items, e.g., would never rank i_4 before i_1. We note that all preference (rating or rank) aggregation strategies are Pareto.

The set of items that are not dominated by any other are called *Pareto optimal.* The top item according to any Pareto aggregation strategy is Pareto optimal. Items i_1 and i_2 comprise the set of Pareto optimal items in the example. Generally, we can define the *N-level Pareto optimal* set to contain items that are dominated by at most $N - 1$ other items. Thus, the top-N choices according to any Pareto aggregation strategy are within the N-level Pareto optimal set. Item i_3 (resp. i_4) is 2-level Pareto optimal as it is dominated by only one item i_2 (resp. i_1). On the other hand items i_5 and i_6 are dominated by at least two and are not 2-level Pareto optimal.

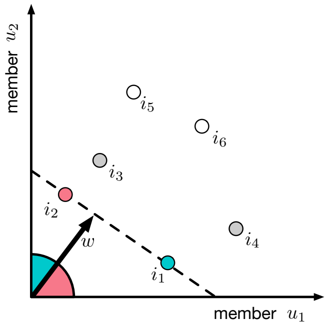


Figure 1: Example of six items depicted as points based on their ranks by the two group members; items i_1 and i_2 are Pareto optimal; items i_1–i_4 are 2-level Pareto optimal.

Approach. Note that to identify the exact set of N-level Pareto optimal items, we would need to obtain the rank of each item to each user. In case this is impractical, we can approximate this set with the following process. We request top-N' recommendations for each group member, where $N' > N$ is the largest number of items the system can recommend, and take their union. Among this subset of items, we identify the N-level Pareto optimal items, which we take to be the approximation to the overall N-level Pareto optimal items. For what follows, we refer to this extracted set as the N-PO items.

Our approach will select N among the PO items to compile the group recommendation list. Particularly, it assigns a score to each PO item and then ranks them decreasingly on this score. We would like the score of an item to represent the probability that any Pareto aggregation strategy would rank the item among the top N choices. As enumerating all possible Pareto aggregation strategies is infeasible, we restrict ourselves to the case of a particular family of aggregation strategies.

Specifically, we consider the class of *linear aggregation strategies*, which includes the average aggregation strategy and each member's individual ranking. A linear strategy assigns weights to each member and essentially computes a weighted sum of the ranks of an item. Therefore, such a strategy can be uniquely represented by its *weight vector* in the $\mathbb{R}^m$ space. Recall, that a hyperplane (line in 2d) orthogonal to the weight vector captures the space of possible items that achieve the same rank. In our example, weight $w = (3/7, 4/7)$ represents the linear aggregation strategy that places i_1 and i_2 at the same rank. As there is no other item that lies on the space between the origin and this line, we deduce that items i_1 and i_2 are at the top rank according to linear strategy w.

In our example with a group of 2 users, observe that it is straightforward to compute the probability with which a linear strategy would rank an item at the top. Only the Pareto optimal items would have a nonzero probability, and, in fact, i_2 has a probability of $4/7$, while i_1 of $3/7$. To understand this, observe that any strategy other than w would either place i_1 or i_2 as the single top choice. A vector assigning a higher relative weight to member u_1 than u_2 would rank i_2 higher than i_1 (i_2 is the best option for u_1 individually, while i_1 is the best for u_2). Any such vector in favor of i_2 has slope that falls in the bottom-right (red shaded) segment of the quadrant, which represents all possible slopes. Conversely, any vector favoring i_1 over i_2 has slope in the top-right (blue shaded) segment. Since the former segment is larger than the latter (by a factor of $4:3$), one can then argue that if all aggregation strategies are equally probable, item i_2 is more likely to be the best choice for the group.

While, such an analytical computation of probabilities of inclusion in the top-N is possible for the case of groups with two members, the general case is not as it entails computing the convex hull of items and enumerating its facets [19]. Thus, we opt for a simple Monte Carlo method for computing said probabilities. Specifically, we generate a large number of random weight vectors, each representing a different linear aggregation strategy, and count how many times each N-PO item ranks within the top-N. Then, items are ranked decreasingly by their counts, and the top-N are returned as the group recommendations. We call this approach N-level Pareto Optimal aggregation, or simply NPO.

Recall than NPO looks for items to recommend among the pool of N-level Pareto optimal items. As this set can be much larger than N, we also consider a variant of NPO that chooses among a smaller pool of items. Specifically, we perform binary search to identify the smallest level $x \in [1, N]$ such that there are at least N items in the x-level Pareto optimal set. Then, we follow the same ranking mechanism based on random weight vectors. We call this approach x-level Pareto Optimal aggregation, or simply XPO.

4 EVALUATION

Section 4.1 describes out experimental setup, while Section 4.2 presents the results of our evaluation.

4.1 Setup

Dataset. For our evaluation, we use the MovieLens 1M dataset [9] containing 6,040 users, 3,952 items (movies), and 1,000,209 ratings. We synthetically create groups of size $m = 2$ up to 8 of two different kinds. In *RND* we assign users to groups sampling uniformly at random from MovieLens. This corresponds to the real life equivalent of groups with unrelated members, such as those visiting a shop. In *SIM* we choose users that are similar to each other, corresponding for example to groups of friends with similar taste. Specifically, starting from a randomly selected user, the group is build incrementally by adding the most similar (in terms of mean-centered cosine similarity) user to the group. Similar to previous work [11, 21, 29], we use a simple matrix factorization technique to fill in the missing ratings in the dataset. For each group, we select 200 items at random, and we set the ground truth of each member to her top-N items among them recommended by the system, where N varies from 5 up to 100.

Methods. Our evaluation compares our methods, denoted as NPO and XPO, with score aggregation, rank aggregation, and fairness-based methods. AVG, LM, MAX, and MUL implement the additive, least-misery, maximum-pleasure, and multiplicative score aggregation strategies, respectively [14]. BORDA and MED are two rank aggregation strategies [2], that assign to an item its average or median rank, respectively; MED is preferred over the Spearman's footrule-based method of [2] as an approximation of the Kemeny optimal rank, due to its lowest time complexity. PEN is the simple method from [10] that introduces a variance-based penalty;

GRF corresponds to the group rating fairness method of [21]; SPG and EPG to the single proportionality and envy-freeness greedy algorithms of [29]; and GVAR to the greedy variance algorithm of [11].

Evaluation Metrics. Each method is evaluated primarily on its fairness and secondarily on its group utility. Both these measures are defined with respect to the member utility which quantifies the similarity of a list σ to the member's ground truth σ_u. We denote by $\sigma[k]$ the item at the k-th position, and by $\sigma[:k]$ the first k items in list σ. We consider four similarity metrics where higher values are better; the first takes values in the range $[-1, 1]$, while all others in $[0, 1]$.

Kendall's τ. A pair of items is concordant if their relative ranking is the same in both lists, and discordant if their relative ranking is reversed in the lists. A pair of items that are at the same rank in one list is neither concordant nor discordant. Any item that does not appear in one list is considered to be at the last rank. Kendall's τ is the number of concordant minus the number of discordant pairs, normalized by the number of possible pairs.

P@k. The precision at position k of a list w.r.t. a member's ground truth measures the ratio of items in the first k positions of the list that appear in the ground truth:

$$\text{P@k} = \frac{|\sigma[1:k] \cap \sigma_u|}{k}.$$

AP. The Average Precision (AP) of a list w.r.t. a member's ground truth is the average of precision at every recall level (a position in the ranking where a relevant item is found), and is computed as:

$$\text{AP} = \frac{1}{|I_g|} \sum_k \text{P@k} \cdot |\sigma[k] \cap \sigma_u|,$$

where $|\sigma[k] \cap \sigma_i|$ indicates whether the item at the k-th position of list σ is in the member's ground truth.

NDCG@k. As in [11], we use Borda semantics and set the relevance of an item at position k in the ground truth equal to $N - k + 1$. We denote the relevance of a ground truth item i as $r_u(i)$. The Discounted Cumulative Gain (DCG) at position k of a list w.r.t. a member's ground truth is:

$$\text{DCG@}k = \sum_{n=1}^{k} \frac{2^{r_u(\sigma[n])}}{\log(n+1)}.$$

IDCG@k is the maximum possible DCG@k, and the Normalized Discounted Cumulative Gain at position k is NDCG@k = DCG@k/IDCG@k.

The reported values of fairness and group utility are the averages among 100 randomly generated groups.

4.2 Evaluation Results

Default Setting. In the first experiment, we investigate the fairness and group utility achieved by all methods, for the default setting where we fix the size of group to $m = 5$ and request the $N = 20$ recommendations. We consider various alternative definitions of member utility: Kendall's τ, average precision, and precision and NDCG at ranks 1, 5, 10, 20.

Table 1 present the results for groups of five random users. The first set of columns correspond to fairness, while the last to group utility. For each column, the best value is shown with bold.

Let us first investigate fairness, where the most important observation is that for all examined flavors, the best value is attained by one of our methods, seven times by XPO and three times by NPO. In all cases, the margin is quite large, e.g., for NDCG@20, XPO achieves 0.579, NPO 0.472, while the third best method, MUL, only 0.438. Moreover, for all flavors except Kendall's τ, the second best method is our other method. Consistently among fairness flavors we find that: among score aggregation techniques, AVG and MUL are the strongest; among rank aggregation, BORDA is the best; among fairness-aware recommenders, PEN and GVAR appear stronger. Across categories, we find that score aggregation techniques perform better than rank aggregation or fairness-aware methods.

The results with respect to group utility are suprisingly analogous. The method that achieves the highest group utility is one of our fairness-aware methods, except in the case of Kendall's τ. The margins are this time smaller, and in some cases the second best method is not our other method, but instead one of AVG, PEN, or GVAR. Regarding our methods, we observe that NPO makes better choices for the first few ranks (higher P@1 and NDCG@1) while XPO makes better choices down the list.

Table 1 present the results for groups of five similar users. The trends shown there are consistent with those in the case of groups with random members. As expected, making recommendations for groups of like-minded users is an easier task and this is evidenced by the higher values across all measures. Again, NPO and XPO are the best two methods for fairness and group utility across all flavors (except Kendall's τ) with a wide margin. Between them, NPO appear stronger when evaluated at the first ranks, but XPO is overall the (τ and AP) the most effective method. Among the other methods, AVG is consistently the third best group recommender.

Varying Group Size. In the second experiment, we investigate the effect of group size in fairness as we request for top $N = 20$ group recommendations; findings on group utility are somewhat similar and omitted. We only show results on member utilities defined as Kendall's τ, AP, NDCG@1, and NDCG@5; other flavors show similar results and are omitted.

Figures 2 and 3 show fairness vs. m for groups of random and similar users, respectively. As the number m of users increases all flavors of fairness decrease, meaning that it becomes harder to provide fair recommendations to more users, especially if they are randomly selected. In all setting tested, the most fair method is one of NPO and XPO. Particularly, XPO achieves a wide margin in terms of fairness measured for the entire recommended list, as shown by member utility defined in terms of τ and AP. NPO is the method that places at the first few ranks items that are considered fair, particularly in the case of smaller groups. For large groups and/or lower ranks, XPO is the most fair. Among the other methods, we observe that in most cases AVG is the most fair with MUL and PEN closely following.

Table 1: Fairness and Group Utility for RND; $m = 5$, $N = 20$

	Fairness										**Group Utility**									
	τ	AP	P@1	P@5	P@10	P@20	ndcg@1	ndcg@5	ndcg@10	ndcg@20	τ	AP	P@1	P@5	P@10	P@20	ndcg@1	ndcg@5	ndcg@10	ndcg@20
AVG	-0.352	0.126	0.490	0.294	0.3	0.267	0.249	0.302	0.314	0.298	-0.221	0.258	0.702	0.575	0.490	0.399	0.694	0.596	0.530	0.455
LM	-0.414	0.071	0.385	0.208	0.196	0.165	0.234	0.226	0.222	0.197	-0.305	0.171	0.632	0.484	0.385	0.276	0.621	0.510	0.431	0.341
MAX	-0.339	0.070	0.424	0.142	0.2	0.223	0.009	0.118	0.177	0.208	**-0.052**	0.224	0.346	0.428	0.424	0.399	0.344	0.409	0.413	0.400
MUL	-0.356	0.121	0.486	0.288	0.295	0.260	0.249	0.299	0.311	0.290	-0.228	0.250	0.702	0.570	0.486	0.389	0.694	0.593	0.526	0.446
BORDA	-0.413	0.075	0.406	0.15	0.206	0.217	0.063	0.136	0.186	0.210	-0.301	0.198	0.478	0.441	0.406	0.352	0.469	0.441	0.417	0.378
MED	-0.435	0.080	0.427	0.1	0.177	0.237	0.008	0.079	0.141	0.211	-0.301	0.225	0.364	0.398	0.427	0.395	0.359	0.383	0.408	0.396
PEN	-0.352	0.123	0.488	0.294	0.298	0.261	0.249	0.304	0.313	0.293	-0.225	0.254	0.702	0.571	0.488	0.394	0.694	0.594	0.528	0.450
GRF	-0.375	0.094	0.473	0.26	0.255	0.226	0.074	0.246	0.258	0.245	-0.231	0.243	0.53	0.550	0.473	0.379	0.521	0.545	0.493	0.423
SPG	-0.434	0.084	0.427	0.1	0.176	0.236	0.063	0.095	0.152	0.218	-0.299	0.230	0.498	0.403	0.427	0.395	0.488	0.411	0.423	0.406
EFG	-0.435	0.082	0.427	0.098	0.177	0.237	0.034	0.087	0.147	0.214	-0.301	0.226	0.404	0.398	0.427	0.395	0.394	0.389	0.411	0.398
GVAR	-0.352	0.123	0.491	0.296	0.299	0.267	0.189	0.294	0.307	0.294	-0.222	0.256	0.664	0.576	0.491	0.399	0.656	0.590	0.526	0.452
NPO	-0.349	0.132	0.482	0.274	0.289	0.281	**0.305**	0.307	0.314	0.311	-0.223	0.264	**0.742**	0.557	0.482	0.415	**0.732**	0.589	0.525	0.466
XPO	**-0.212**	**0.205**	**0.566**	**0.386**	**0.396**	**0.376**	0.268	**0.383**	**0.403**	**0.406**	-0.094	**0.326**	0.724	**0.629**	**0.566**	**0.485**	0.716	**0.642**	**0.594**	**0.532**

Table 2: Fairness and Group Utility for SIM; $m = 5$, $N = 20$

	Fairness										**Group Utility**									
	τ	AP	P@1	P@5	P@10	P@20	ndcg@1	ndcg@5	ndcg@10	ndcg@20	τ	AP	P@1	P@5	P@10	P@20	ndcg@1	ndcg@5	ndcg@10	ndcg@20
AVG	-0.241	0.239	0.633	0.552	0.438	0.367	0.816	0.609	0.510	0.437	-0.091	0.390	0.924	0.755	0.633	0.495	0.923	0.793	0.695	0.577
LM	-0.262	0.211	0.590	0.54	0.388	0.327	0.816	0.602	0.476	0.403	-0.120	0.354	0.924	0.745	0.590	0.452	0.923	0.785	0.664	0.542
MAX	-0.187	0.151	0.559	0.276	0.326	0.332	0.135	0.252	0.313	0.329	**0.149**	0.345	0.496	0.556	0.559	0.526	0.494	0.543	0.549	0.530
MUL	-0.242	0.238	0.632	0.556	0.434	0.366	0.816	0.612	0.508	0.438	-0.091	0.389	0.924	0.758	0.632	0.494	0.923	0.795	0.694	0.576
BORDA	-0.458	0.085	0.341	0.096	0.148	0.235	0.077	0.101	0.134	0.203	-0.358	0.206	0.304	0.318	0.341	0.371	0.301	0.315	0.332	0.357
MED	-0.527	0.066	0.275	0.056	0.11	0.185	0.020	0.045	0.090	0.154	-0.420	0.182	0.196	0.229	0.275	0.339	0.196	0.217	0.252	0.306
PEN	-0.242	0.237	0.632	0.552	0.434	0.364	0.816	0.609	0.507	0.437	-0.092	0.388	0.924	0.754	0.632	0.493	0.923	0.792	0.694	0.575
GRF	-0.329	0.150	0.542	0.292	0.332	0.338	0.137	0.259	0.301	0.321	-0.183	0.317	0.368	0.506	0.542	0.484	0.367	0.475	0.511	0.483
SPG	-0.525	0.069	0.276	0.056	0.11	0.185	0.097	0.060	0.099	0.159	-0.417	0.187	0.332	0.234	0.276	0.339	0.329	0.246	0.269	0.317
EFG	-0.527	0.067	0.275	0.056	0.11	0.185	0.020	0.049	0.092	0.155	-0.420	0.182	0.2	0.229	0.275	0.339	0.199	0.218	0.252	0.306
GVAR	-0.250	0.206	0.630	0.512	0.438	0.367	0.176	0.472	0.435	0.390	-0.105	0.357	0.4	0.710	0.630	0.496	0.399	0.666	0.626	0.534
NPO	-0.234	0.268	0.650	**0.636**	0.46	0.377	**0.922**	**0.708**	0.561	0.472	-0.103	0.413	**0.988**	**0.822**	0.650	0.509	**0.984**	**0.862**	0.728	0.601
XPO	**-0.003**	**0.377**	**0.739**	0.616	**0.582**	**0.508**	0.855	0.668	**0.632**	**0.579**	0.111	**0.495**	0.948	0.801	**0.739**	**0.605**	0.947	0.832	**0.779**	**0.674**

Varying Top-N. In the last experiment, we fix the group size to $m = 5$ and request from $N = 5$ up to 100 group recommendations. As before, we report the fairness flavors on Kendall's τ, AP, NDCG@1, and NDCG@5, and omit results on group fairness.

Figures 4 and 5 depict fairness vs. N for groups of random and similar users, respectively. Increasing the number of requested recommendations means that the length of the ground truth lists also increases, and thus all flavors of fairness increase with N. In almost all settings, NPO or XPO are the most fair group recommenders, but their margin decreases with N. It is worth mentioning that for groups of similar users when $N = 100$, the most fair method in terms of NDCG@1 and NDCG@5 is plain AVG. In this setting, half of the items to choose from are relevant, and thus NDCG values approach 1. For smaller values of N, and particularly in terms of NDCG-based fairness, our methods offer significant benefits in groups of similar users.

5 CONCLUSIONS

In this work, we propose two novel top-N group recommenders that are fairness-aware. More precisely, the notion of fairness is intuitively defined in terms of the minimum utility any group member attains from the recommended list. A member's utility is in turn relative to her own's best possible list of recommendations. The main idea behind our recommenders is that they focus on the Pareto optimal items, and seek to rank them in an objective and fair manner that treats each member equally. Specifically, our methods consider a class of Pareto efficient aggregation strategies and estimate the probability of an item to belong within the top-N ranks of any such strategy. Items are then ranked decreasingly by their estimated probability.

Experiments on synthetically generated groups over a real dataset show that in a variety of settings our two group recommenders are not only the best in terms of fairness, but also in terms of the overall group utility. Our study has also made some interesting observation regarding existing group recommenders, fairness-aware or not. The effectiveness of all methods decreases as the number of random users in a group increases. The phenomenon is much less pronounced in groups of similar users. In general, high inter group similarity tends to favor non fairness-based methods. Surprisingly, we find that score aggregation methods are more fair than existing fairness-aware methods, with the traditional average and multiplicative strategies being the best performers.

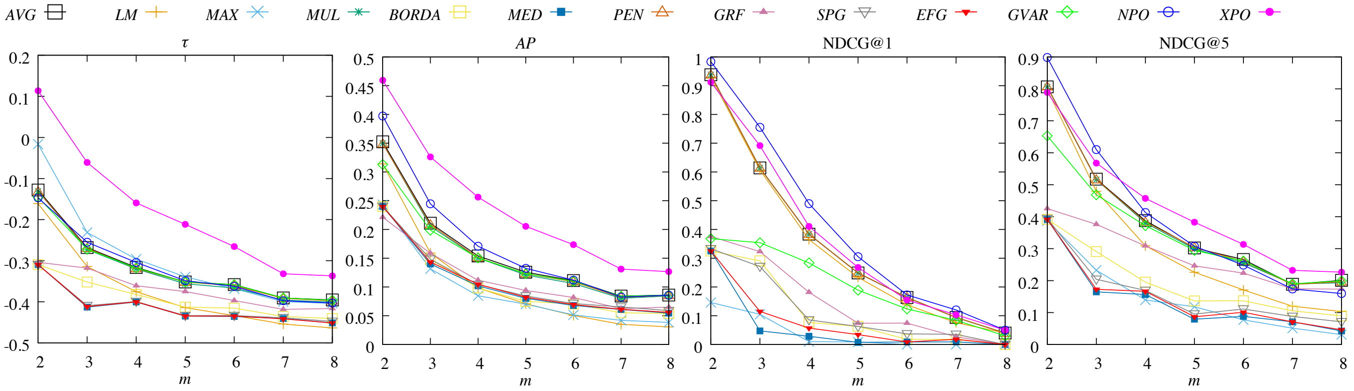
In the future, we would like to closely investigate the performance of our two methods, and identify exactly when and why one is expected to outperform the other. The goal is to define a hybrid that is the single most fair method across settings.

REFERENCES

[1] Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. 2003. Intrigue: Personalized Recommendation of Tourist Attractions for Desktop and Hand Held Devices. *Applied Artificial Intelligence* 17, 8-9 (2003), 687–714. https://doi.org/10.1080/713827254

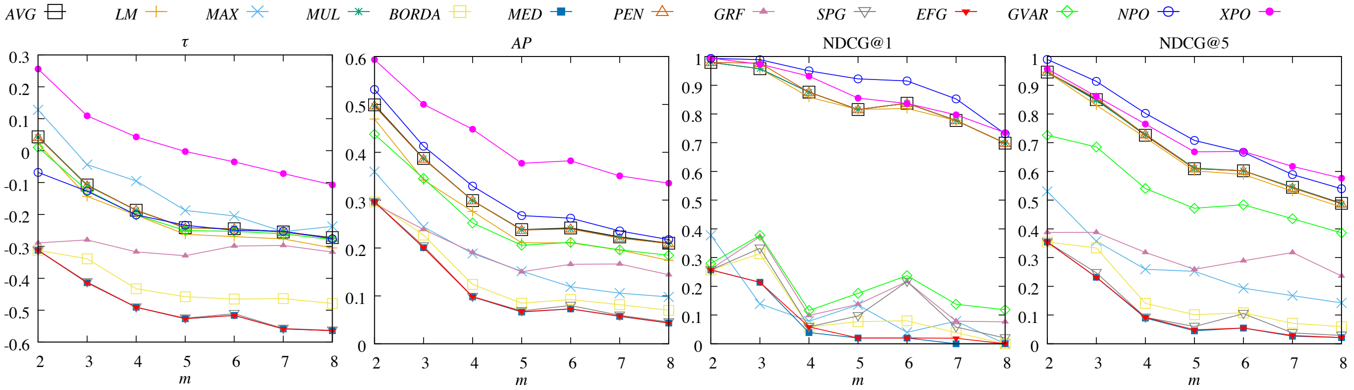
[2] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. 2010. Group recommendations with rank aggregation and collaborative filtering. In *RecSys*. 119–126. https://doi.org/10.1145/1864708.1864733

[3] Shlomo Berkovsky and Jill Freyne. 2010. Group-based recipe recommendations: analysis of data aggregation strategies. In *RecSys*. 111–118. https://doi.org/10.1145/1864708.1864732



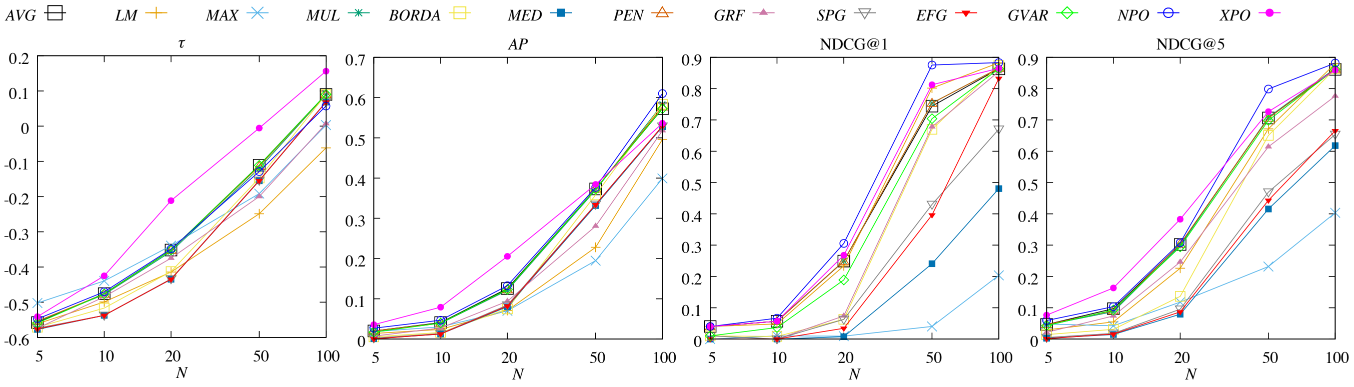
(a) Kendall's τ **(b) Average Precision** **(c) NDCG@1** **(d) NDCG@5**

Figure 2: Fairness vs. group size m for RND; $N = 20$



(a) Kendall's τ **(b) Average Precision** **(c) NDCG@1** **(d) NDCG@5**

Figure 3: Fairness vs. group size m for SIM; $N = 20$



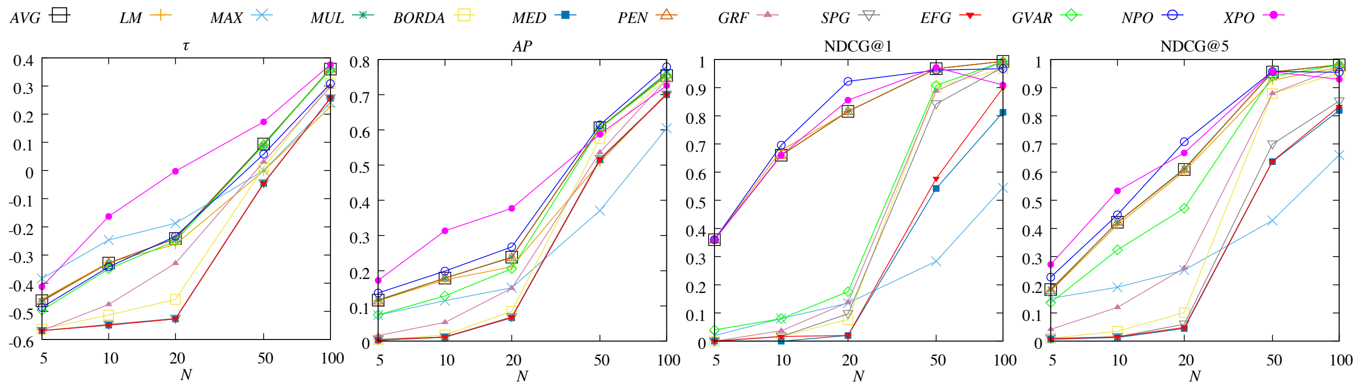
(a) Kendall's τ **(b) Average Precision** **(c) NDCG@1** **(d) NDCG@5**

Figure 4: Fairness vs. top-N for RND; $m = 5$

[4] Robin Burke. 2017. Multisided Fairness for Recommendation. *CoRR* abs/1707.00093 (2017). arXiv:1707.00093 http://arxiv.org/abs/1707.00093

[5] Yen-Liang Chen, Li-Chen Cheng, and Ching-Nan Chuang. 2008. A group recommendation system with consideration of interactions among group members. *Expert Syst. Appl.* 34, 3 (2008), 2082–2090. https://doi.org/10.1016/j.eswa.2007.02.008

[6] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. 2001. Rank aggregation methods for the Web. In *WWW*. 613–622. https://doi.org/10.1145/371920.372165



(a) Kendall's τ **(b) Average Precision** **(c) NDCG@1** **(d) NDCG@5**

Figure 5: Fairness vs. top-N for SIM; $m = 5$

[7] Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Comparing Top k Lists. *SIAM J. Discrete Math.* 17, 1 (2003), 134–160. http://epubs.siam.org/sam-bin/dbq/article/41285

[8] Mike Gartrell, Xinyu Xing, Qin Lv, Aaron Beach, Richard Han, Shivakant Mishra, and Karim Seada. 2010. Enhancing group recommendation by incorporating social relationship interactions. In *Proceedings of the 2010 International ACM SIGGROUP Conference on Supporting Group Work, GROUP 2010, Sanibel Island, Florida, USA, November 6-10, 2010*, Wayne G. Lutters, Diane H. Sonnenwald, Tom Gross, and Madhu Reddy (Eds.). ACM, 97–106. https://doi.org/10.1145/1880071.1880087

[9] F. Maxwell Harper and Joseph A. Konstan. 2016. The MovieLens Datasets: History and Context. *TiiS* 5, 4 (2016), 19:1–19:19. https://doi.org/10.1145/2827872

[10] Anthony Jameson and Barry Smyth. 2007. Recommendation to Groups. In *The Adaptive Web, Methods and Strategies of Web Personalization*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Springer, 596–627.

[11] 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*, Paolo Cremonesi, Francesco Ricci, Shlomo Berkovsky, and Alexander Tuzhilin (Eds.). ACM, 107–115. https://doi.org/10.1145/3109859.3109887

[12] Xingjie Liu, Yuan Tian, Mao Ye, and Wang-Chien Lee. 2012. Exploring personal impact for group recommendation. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki (Eds.). ACM, 674–683. https://doi.org/10.1145/2396761.2396848

[13] Judith Masthoff. 2004. Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. *User Model. User-Adapt. Interact.* 14, 1 (2004), 37–85. https://doi.org/10.1023/B:USER.0000010138.79319.fd

[14] Judith Masthoff. 2015. Group Recommender Systems: Aggregation, Satisfaction and Group Attributes. In *Recommender Systems Handbook*. 743–776. https://doi.org/10.1007/978-1-4899-7637-6_22

[15] Judith Masthoff and Albert Gatt. 2006. In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems. *User Model. User-Adapt. Interact.* 16, 3-4 (2006), 281–319. https://doi.org/10.1007/s11257-006-9008-3

[16] Joseph F. McCarthy and Theodore D. Anagnost. 1998. MusicFX: An Arbiter of Group Preferences for Computer Supported Collaborative Workouts. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW '98)*. ACM, New York, NY, USA, 363–372. https://doi.org/10.1145/289444.289511

[17] Mark O'Connor, Dan Cosley, Joseph A. Konstan, and John Riedl. 2001. PolyLens: A recommender system for groups of user. In *ECSCW*. 199–218.

[18] Toon De Pessemier, Simon Dooms, and Luc Martens. 2014. Comparison of group recommendation algorithms. *Multimedia Tools Appl.* 72, 3 (2014), 2497–2541. https://doi.org/10.1007/s11042-013-1563-0

[19] Franco P. Preparata and Michael Ian Shamos. 1985. *Computational Geometry: An Introduction.* Springer.

[20] Shuyao Qi, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2016. Recommending Packages to Groups. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, Francesco Bonchi, Josep Domingo-Ferrer, Ricardo A. Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu (Eds.). IEEE, 449–458. https://doi.org/10.1109/ICDM.2016.0056

[21] Shuyao Qi, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2018. Recommending packages with validity constraints to groups of users. *Knowl. Inf. Syst.* 54, 2 (2018), 345–374. https://doi.org/10.1007/s10115-017-1082-9

[22] Juan A. Recio-García, Guillermo Jiménez-Díaz, Antonio A. Sánchez-Ruiz-Granados, and Belén Díaz-Agudo. 2009. Personality aware recommendations to groups. In *Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009, New York, NY, USA, October 23-25, 2009*, Lawrence D. Bergman, Alexander Tuzhilin, Robin D. Burke, Alexander Felfernig, and Lars Schmidt-Thieme (Eds.). ACM, 325–328. https://doi.org/10.1145/1639714.1639779

[23] Dimitris Sacharidis. 2017. Group Recommendations by Learning Rating Behavior. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP 2017, Bratislava, Slovakia, July 09 - 12, 2017*, Mária Bieliková, Eelco Herder, Federica Cena, and Michel C. Desmarais (Eds.). ACM, 174–182. https://doi.org/10.1145/3079628.3079691

[24] Amirali Salehi-Abari and Craig Boutilier. 2015. Preference-oriented Social Networks: Group Recommendation and Inference. In *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015*, Hannes Werthner, Markus Zanker, Jennifer Golbeck, and Giovanni Semeraro (Eds.). ACM, 35–42. https://doi.org/10.1145/2792838.2800190

[25] Lara Quijano Sánchez, Juan A. Recio-García, Belén Díaz-Agudo, and Guillermo Jiménez-Díaz. 2013. Social factors in group recommender systems. *ACM TIST* 4, 1 (2013), 8:1–8:30. https://doi.org/10.1145/2414425.2414433

[26] Shunichi Seko, Takashi Yagi, Manabu Motegi, and Shin-yo Muto. 2011. Group recommendation using feature space representing behavioral tendency and power balance among members. In *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius (Eds.). ACM, 101–108. https://doi.org/10.1145/2043932.2043953

[27] Christophe Senot, Dimitre Kostadinov, Makram Bouzid, Jérôme Picault, and Armen Aghasaryan. 2011. Evaluation of Group Profiling Strategies. In *IJCAI*. 2728–2733. http://ijcai.org/papers11/Papers/IJCAI11-454.pdf

[28] Christophe Senot, Dimitre Kostadinov, Makram Bouzid, Jérôme Picault, Armen Aghasaryan, and Cédric Bernier. 2010. Analysis of Strategies for Building Group Profiles. In *User Modeling, Adaptation, and Personalization, 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010. Proceedings (Lecture Notes in Computer Science)*, Paul De Bra, Alfred Kobsa, and David N. Chin (Eds.), Vol. 6075. Springer, 40–51. https://doi.org/10.1007/978-3-642-13470-8_6

[29] 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*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 371–379. https://doi.org/10.1145/3038912.3052612

[30] Martin Stettinger. 2014. Choicla: towards domain-independent decision support for groups of users. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren (Eds.). ACM, 425–428. https://doi.org/10.1145/2645710.2653365

[31] Peyton Young. 1995. Optimal voting rules. *Journal of Economic Perspectives* 9, 1 (1995), 51–64.

[32] Zhiwen Yu, Xingshe Zhou, Yanbin Hao, and Jianhua Gu. 2006. TV Program Recommendation for Multiple Viewers Based on user Profile Merging. *User Model. User-Adapt. Interact.* 16, 1 (2006), 63–82. https://doi.org/10.1007/s11257-006-9005-6