Coupled or Decoupled Evaluation for Group Recommendation Methods?

Ladislav Peska[1], Ladislav Malecek[1]

[1]Charles University, Faculty of Mathematics and Physics, Malostranské nám. 25, Prague, Czech Republic

Abstract

Group recommendations are a sub-domain of recommender systems (RS), where the final recommendations should comply with preferences of all members of the group. Usually, group recommendations are built on top of common "single-user" RS via aggregating models or predictions for multiple users with some notions of fairness and relevance in mind.

So far, group recommendations were usually evaluated off-line either as a tightly coupled pair with the underlying RS or in a decoupled fashion, where the relevance scores estimated by underlying RS serves as a ground truth. Both evaluation types may suffer from different biases that provide illicit advantages to some classes of group recommending strategies. In experimental part, we evaluate several recent group recommendation models and show that the evaluation process itself significantly affects their perceived usability. While coupled evaluation favors group RS that tend to select per-user best items, decoupled evaluation favors strategies aiming to find items with (some degree of) overall agreement. We further evaluate methods w.r.t several variants of inverse propensity based de-biasing scenario in order to reduce the popularity bias of coupled evaluations. Also in this case, if groups of similar users are considered, the magnitude of de-biasing has a determining effect on the ordering of individual methods.

Keywords

Group recommender systems, Popularity Bias, Evaluation protocols

1. Introduction and Related Work

Group recommender systems [1] are an interesting sub-area of recommender systems (RS) research. Instead of focusing on the preferences of an individual, group recommendations should be provided in accordance to the preference of several users. Although there are not many reported usages of group RS in praxis yet[1], several domains could benefit from group RS, e.g. movies or music streaming services. Groups of people often gather together to watch a movie, while events such as parties or shared car rides often involves listening to music. It seems desirable that in such cases, preferences of all participants would be reflected to some extent. Therefore it is only natural to expect similar behavior from the recommender systems targeting such groups. In another words, recommendations should be to some extent fair for the participants.

Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2021), September 25th, 2021, co-located with the 15th ACM Conference on Recommender Systems, Amsterdam, The Netherlands

peska@ksi.mff.cuni.cz (L. Peska); malecek.ladislav@gmail.com (L. Malecek)

0000-0001-8082-4509 (L. Peska)

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

[1]Some exceptions are e.g. FlyTrap [2], Pocket RestaurantFinder [3] or HappyMovie [4].

In this paper, we focus on various fairness-preserving group recommendation strategies that operate on top of classical (single user) recommending systems (e.g. [5, 6, 7, 8]). We will further denote them as *group RS aggregators*. On its input, group RS aggregators expect individual recommendations for each of the group members (possibly with additional metadata) and they return the final list of recommendations for the whole group. Utility functions based on which the final recommendations are derived usually consider some notions of per-user fairness as well as overall per-item relevance. Although there are some exceptions (e.g. [8, 9, 10]), vast majority of approaches treat individual group members uniformly and do not consider e.g. power relationships involved in groups, opinion leaders or long-term effects. As group dynamics are not in the centre of this work, we followed the uniform treatment of users utilized in the majority of surveyed papers.

We can roughly divide group RS aggregators to *item-wise* and *list-wise*. *Item-wise* approaches evaluate the utility function for each item separately, irrespective of other items' scores. Many utility functions were proposed in the history, e.g. *Least Misery*, *Average*, *Average without Misery*, *Borda count* and many more. Item-wise approaches were thoroughly surveyed in [1].

List-wise approaches argue that item-wise approaches may introduce a systematic bias against some group members, e.g. if his/her preferences differs from the rest of the group [5, 8]. Instead, they usually construct the final list of recommendations incrementally, while the context of previously selected recommendations is considered when the next item is being selected. An early example is FAI algorithm [1], which regularly switches between users and selects the best remaining item of the current user. Some of the recently published list-wise approaches are SPGreedy [6], GreedyLM [7], XPO [11], GFAR [5] or EP-FuzzDA [8].

The main concern of this paper is the evaluation of group RS aggregators. Having a couple of underlying (single user) recommender and a group RS aggregator leaves two principal options to conduct the evaluation. First, one can perform a *coupled* evaluation, i.e. consider the RS and the group aggregator as a tightly coupled pair and evaluate their overall performance as in [7, 5]. Another option is a *decoupled* evaluation, where authors aim to evaluate the performance of group RS aggregators themselves and ratings/ranking provided by the underlying RS is considered as a ground truth [6, 11, 8].

In this paper, we show that both evaluation types may introduce certain biases and provide considerably different results. Furthermore, in order to provide bias-free estimation of group RS aggregator's performance, we utilized coupled evaluation with *Self-normalized Inverse Propensity Score (SNIPS)* evaluator[12]. Notably, not only the results of coupled and decoupled evaluations differ significantly, but the relative performance of individual aggregators also greatly depend on the magnitude of de-biasing when SNIPS evaluator is employed.

We will continue with a brief description of evaluation protocols followed by the results presentation and a discussion on possible causes and implications of the observed results.

2. Evaluation Protocols for Group RS Aggregators

Evaluation protocols of recently proposed group RS aggregators differed considerably. Therefore, in this section, we provide a brief overview of utilized variants.

2.1. Preliminaries

The first design choice of every evaluation protocol is the selection of underlying RS. While the choice of RS is in theory orthogonal to the rest of the evaluation protocol, it may have some interesting implications for both coupled and decoupled evaluation protocols. Most related papers ([7, 11, 5, 8]) utilized some variant of matrix factorization, e.g. ALS [13]. One exception was the work of Serbos et al. [6], who used Item-based KNN [14]. This may be relevant distinction if the decoupled evaluation is considered as item-based KNN cannot predict preference for all user-item pairs.

Next, the groups of users whom the recommendations are to be addressed have to be assembled. Finding a suitable dataset containing actual groups of users is a prevalent problem. Xiao et al. [7] utilized MoviePilot dataset, where a fraction of users have shared accounts, i.e., groups. Nonetheless the volume of such groups is rather low and they mostly have only two members, which limits its applicability.

Other than this, authors resort to artificially constructed groups on standard datasets. One can either use random users, or their similarity is considered. For instance, Kaya et al. [5] considered random, similar and divergent user groups. Similar and divergent groups were constructed iteratively w.r.t. pairwise correlation of users' ratings.[2] Nonetheless, authors claimed that performance w.r.t. divergent and random groups was highly similar, so in this paper, we adopted the similar and divergent group definitions from [5].

2.2. Coupled Evaluation

Coupled evaluation protocol (see Figure 1 top) closely resembles the standard static off-line evaluation of RS (with imputed group RS aggregator). Historical user feedback ($r_{u,i} \in \mathcal{R}$ is first divided into train set and test set (cross-validation is often applied) and train set is forwarded to the RS. Recommender system outputs estimated preferences ($\hat{r}_{u,i}$) of individual users and push them to the group RS aggregator, which provides final recommendations for the whole group (list L_G of top-20 recommendations in our case). Performance of the overall solution is evaluated w.r.t. withheld fraction of the user feedback. As such, this evaluation strategy estimates the performance of RS and group aggregator *couple*. Kaya et al. [5] and Xiao et al. [7] utilized coupled evaluation.

The obvious disadvantage of this evaluation strategy is its inability to completely disentangle the performance of group aggregators from the performance of underlying RS (i.e., the way how RS recommend may be more/less suitable for individual group RS aggregators). This problem can be to some extent solved by utilizing multiple, sufficiently diverse, recommender systems.

The second disadvantage is that in vast majority of datasets, only a small fraction of potentially relevant items is known and the missing feedback is not randomly distributed. Feedback on relevant long-tail items is missing more often than feedback on highly popular items (i.e., missing-not-at-random problem, MNAR [12]). Furthermore, collaborative filtering RS such as variants of matrix factorization tends to exhibit popularity bias [15]. I.e., popular items tend to be recommended more frequently (and on higher positions) than what would be proportional to

[2] To be more specific, next group member was selected at random from users who have Pearson's correlation of at least 0.3 (similar) or no more than 0.1 (divergent) to some of the existing group members.

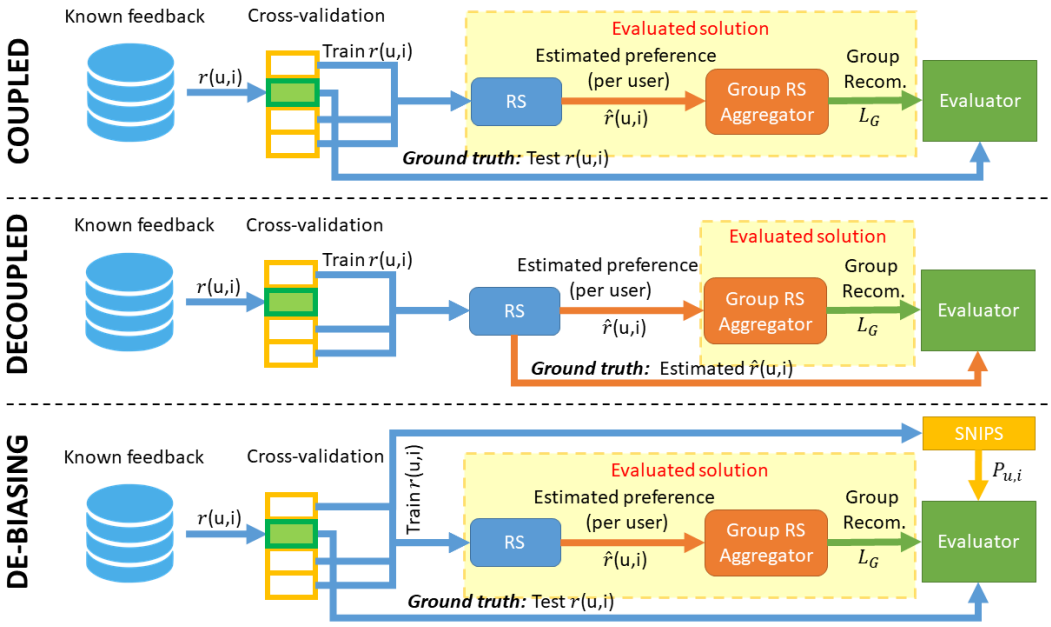

Figure 1: Illustratory examples of three evaluation protocols: *coupled*, *decoupled* and *de-biasing*. The main difference between coupled and decoupled evaluation is the source of ground truth, while de-biasing protocol extends coupled protocol by inverse propensity based results normalization.

their overall popularity. Combined with the MNAR problem and coupled evaluation, popularity biased RS may provide an illicit favor for such group recommendation strategies that often select top recommended items for each user (such as FAI or GFAR) over strategies recommending items that are acceptable (not necessarily best) for the majority of users.

2.3. Decoupled Evaluation

In decoupled evaluation, estimated ratings $\hat{r}_{u,i}$ received from the RS are considered as the ground truth (see Figure 1 middle). With this, we simulate the situation where all user preferences are known and we only evaluate the capability of group RS aggregators to effectively combine those preferences. Malecek et al. [8], Serbos et al. [6] and Sacharidis [11] utilized decoupled evaluation. Nonetheless, we need to note that in [11] only a randomly selected subset of items were evaluated and item-based KNN utilized in [6] was not able to derive estimated preferences to some user-item pairs. These differences can be seen as modifications to the decoupled evaluation protocol.

As in coupled evaluation, a possible problem of this evaluation scenario is the interference between underlying RS and group aggregation strategies (i.e., the general characteristics of supplied ratings may be more suitable for some approaches than others). For instance, consider a RS that systematically overestimates the true relevance of items. In this case, many items that are actually not preferred would be rendered as (mildly) preferred by users. Approaches seeking overall agreement on items could illicitly benefit from this bias, because they receive a broader space of possible combinations to choose from.

2.4. De-biasing Coupled Evaluation via Inverse Propensity Score

Both of the previously described evaluation approaches may introduce a bias that would provide an illicit advantage to some class of group RS aggregators. In order to compare the extent of these biases as well as to mitigate some of them, we considered the utilization of a de-biasing evaluation strategy. Specifically, we focused on de-biasing the popularity bias in coupled evaluation strategies. In order to do so, we utilized the self-normalized inverse propensity score (SNIPS) approach proposed by Yang et al. [12] (further denoted as *de-biasing evaluation strategy*). De-biasing evaluation strategy is essentially a coupled one, but the SNIPS score is utilized as a normalization to reduce an impact of items with high overall popularity (see Figure 1 bottom).

For the set of user's known relevant items (R_u), estimated item's propensity score $P_{u,i}$ and some scoring metric $score(L_G, i)$, the relevance of the list of recommendations L_G for the user u is calculated as follows:

$$r_{L_G,u}^{SNIPS} = \frac{1}{\sum_{i \in R_u} \frac{1}{P_{u,i}}} \sum_{i \in R_u} \frac{score(L_G, i)}{P_{u,i}} \tag{1}$$

For the two metrics considered in this paper, average relevance and discounted cumulative gain at top-20 recommendations, the corresponding scoring functions are as follows:

$$\begin{aligned} score_{AR}(L_G, i) &= \begin{cases} r_{u,i}/|L_G| & \text{if } i \in L_G \\ 0 & \text{otherwise} \end{cases} \\ score_{DCG}(L_G, i) &= \begin{cases} r_{u,i}/log_2(rank(i, L_G)) & \text{if } i \in L_G \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{2}$$

Nonetheless, the question is how to estimate the propensity score $P_{u,i}$. Yang et al. decomposed $P_{u,i}$ to the probability that item is recommended by RS and the probability that item is interacted with if recommended. Authors further assume user independence, completeness of user's observation and popularity bias for the probability of being recommended and finally propose the following estimation of the propensity score:

$$P_{u,i} \propto pop_i^{\frac{\gamma+1}{2}} \tag{3}$$

where pop_i denotes observed popularity of an item i (i.e., the volume of known interactions with item i) and power-law exponent γ is a hyperparameter of the model affecting the propensity distributions over items with various observed popularity levels. Larger γ values lead to lower propensity scores for long-tail items and higher scores for popular ones.

Our research question is whether the popularity bias indeed affects the results of individual group RS strategies as assumed. In order to answer it, we manipulate the γ hyperparameter to gradually decrease the effect of popularity bias in evaluations. We hypothesize that by doing this, the de-biasing evaluation results should gradually resemble the results obtained from decoupled evaluation (i.e., that the de-biasing evaluation strategy can serve as a smooth bridge between coupled and decoupled evaluation strategies). Obviously, the transition cannot be perfect, because no counter-measures were applied for the possible biases introduced by the decoupled evaluation protocol.

3. Comparison of Evaluation Protocols

In order to corroborate our theories on biases introduced by coupled and decoupled evaluation strategies and their impact on various classes of group RS aggregators, we conducted a following experiment. Eight group RS aggregation strategies were evaluated w.r.t. coupled, decoupled and a range of de-biasing evaluation protocols. By observing relative differences of per-approach results w.r.t. individual evaluation protocols, we can derive conclusions on the presence and effects of said biases. Let us first describe the evaluated group RS strategies and parameters of the experiment.

3.1. Group RS Aggregators

In the evaluation, we considered one item-wise and seven list-wise group RS aggregation strategies. Their brief description and references follows. In description, we mainly focused on the estimation, whether the considered algorithm is more biased towards finding an overall agreement between users or whether it rather tend to provide a selection of (close to) best items for individual users.

AVG [1] is a simple item-wise approach that for each item evaluates its mean per-user relevance $r_i = \sum_{u \in G} \hat{r}_{u,i}/|G|$ and returns *top-k* items with the highest relevance. As such, AVG would prefer overall good items, however it does not compensate for any systematic biases. For example, if two group members are highly similar and diverse to the third group member, recommendations suitable for the third member would be missing in the final list.

FAI [1] is a well known list-wise approach that iterates over group members and each time selects the best remaining item for the current group member. As such, FAI exploits the ordering given by the underlying RS, but does not try to find items that are simultaneously suitable for multiple users.

SPGreedy [6] is a list-wise iterative approach that considers so called m-proportionality fairness metric. The list L_G is m-proportional for a user u if at least m items are within the *top-k* best items for user u. SPGreedy algorithm iteratively selects items whose inclusion would maximize the fraction of users for which the L_G is 1-proportional. The behavior of SPGreedy is highly dependent on the considered size of *top-k*. For larger values, items with higher overall agreement would be recommended, while smaller *top-k* would result in more often selection of per-user best items. As in [5], we utilized $k = 1$, which renders SPGreedy among algorithms preferring per-user best items.

GreedyLM [7] approach considers sums of the per-user relevances $\hat{r}_u = \sum_{i \in L_G} \hat{r}_{u,i}$ of the partially constructed L_G list. Specifically, authors employed the *least misery* fairness metric ($LM = min_{u \in G} \hat{r}_u$) and iteratively select the next item based on the linear combination of item's mean utility $r_i = \sum_{u \in G} \hat{r}_{u,i}$ and least misery fairness of the produced list ($\hat{r}_u[L_G \cup \{i\}]$).

XPO [11] considers the concept of Pareto domination w.r.t. per-user ranking[3]. L_G list of the size N is generated from items that are dominated by at most x items, where x is set as smallest number such that the size of the resulting set of candidates is at least N. Finally, L_G is selected via probabilistic approach. A series of weighted averages of per-user ranks is generated (with weights selected at random). For each item, it is calculated how many times it fits into

[3] I.e. item i_1 dominates item i_2 if it is better for at least one user and better or equal for all others.

top-N best items and items with the highest counts are recommended. Given the pre-selection procedure that minimize the volume of eligible candidates, XPO would tend to propose per-user best items.

GFAR [5] defines the fairness through the sum of probabilities that at least one recommended item is relevant for the user: $f_{GFAR}(L_G) = \sum_{u \in G} p(rel|u, L_G)$, which is expressed as the complement to the probability that all items are irrelevant $p(rel|u, L_G) = 1 - \prod_{i \in L_G}(1 - p(rel|u, i))$. Relevance probabilities of individual items $p(rel|u, i)$ are defined as the normalized Borda-count induced by the individual preferences of user u. GFAR utilizes a greedy approach to iteratively construct L_G list. Due to the choice of rather drastic relevance probability estimation as well as the assumption that single relevant item per user is sufficiently fair, GFAR tend to select rather per-user best items than those with (certain level of) overall agreement.

FuzzDA [8] approach is based on D'Hondt's mandates allocation strategy, but extends it with fuzzy candidate-party membership (in the context of group RS aggregation, it allows that an item is preferred by multiple users to a certain degree). Each group member receives certain amount of initial votes v_u and at each step the item with the highest weighted relevance $r_i = \sum_{u \in G} \hat{r}_{u,i} * \bar{v_u}$ is selected. Current per-user votes $\bar{v_u}$ are reduced proportionally to the user-item relevance of currently selected item.

EP-FuzzDA [8] approach is based on FuzzDA, but modifies the item selection procedure. For each user and each iteration, EP-FuzzDA calculates the amount of relevance that is missing ($r_{missing,u}$) to have exactly proportional representation of user's votes in the partially constructed L_G. Then the item with highest (constrained) overall relevance $r_i = \sum_{u \in G} min(\hat{r}_{u,i}, r_{missing,u})$ is selected. Both FuzzDA and EP-FuzzDA tend to select items with higher overall agreement rather than best items for individual users.

3.2. Datasets and Evaluation Details

The experimental setup utilized in this paper is partially based on [5] (datasets, group definitions and underlying RS were the same). We used two datasets from the domains with a potential to utilize group recommendations: movies (ML1M [16]) and music (KGRec music dataset [17]). We utilized 5-fold cross-validation with 60% train, 20% validation and 20% test sets (test set was not used in decoupled evaluation). Estimated per-user relevance scores were supplied by the ALS matrix factorization algorithm [13]. We considered synthetic groups of users of two kinds: with *similar* members and with *divergent* members according to user's rating patterns. For each group type, group sizes from $s = 2$ to $s = 8$ were considered, while up to 1000 synthetic groups were generated for each combination of group size and type. During evaluation, estimated user-item relevance score matrix is calculated by ALS MF for each fold and forwarded to the group recommendation strategies. Finally, each strategy produces top-20 items recommended for the group.

Coupled and de-biasing evaluation scenarios utilize the test set data. KGRec dataset contains binary interactions, but ML1M dataset contains graded relevance feedback (1*-5* ratings). In accordance with [5], we binarized this feedback to resemble expected consumption behavior. To be more specific, in evaluations of ML1M we considered only 4* and 5* ratings to be positive, i.e. $r_{u,i} = 1$ and all others as unknown or negative, i.e. $r_{u,i} = 0$. Decoupled scenario was evaluated w.r.t. estimated per-user relevance scores $\hat{r}_{u,i}$ as supplied by ALS matrix factorization with no

Table 1
Results of coupled and decoupled evaluation of ML1M dataset (top) and KGRec dataset (bottom). For the sake of space we only depict the results of *similar* groups with group size $s = 8$. mean, min and M/M stands for the mean score per group, minimal score per group and ratio between minimal and maximal score per group respectively. Averaged results for all groups are displayed. Best results are in bold, second best are underlined and third best are in italic.

a) ML1M, sim, $s = 8$	**coupled**; AR			nDCG			**decoupled**; AR			nDCG		
	mean	min	M/M	mean	min	M/M	mean	min	M/M	mean	min	M/M
AVG [1]	0.133	0.020	0.066	0.175	0.022	0.058	**2.025**	*1.282*	0.462	**0.569**	*0.402*	*0.565*
FAI [1]	0.133	*0.036*	**0.139**	0.176	*0.039*	**0.119**	1.546	1.059	*0.515*	0.430	0.278	0.478
SPGreedy [6]	0.138	0.029	0.097	*0.186*	0.034	0.092	1.905	1.272	0.498	0.518	0.367	0.559
GreedyLM [7]	*0.140*	0.029	0.096	0.183	0.032	0.086	*1.918*	1.336	0.530	*0.544*	0.419	0.637
XPO [11]	**0.147**	0.039	*0.133*	0.195	0.043	*0.115*	1.712	1.165	0.516	0.493	0.313	0.477
GFAR [5]	0.147	**0.040**	0.137	**0.203**	**0.045**	0.116	1.705	1.148	0.509	0.497	0.304	0.453
FuzzDA [8]	0.116	0.017	0.057	0.154	0.019	0.049	1.833	1.136	0.428	0.517	0.355	0.513
EP-FuzzDA [8]	0.119	0.024	0.094	0.157	0.026	0.077	1.940	**1.609**	**0.730**	0.556	**0.459**	**0.709**

a) KGRec, sim, $s = 8$	**coupled**; AR			nDCG			**decoupled**; AR			nDCG		
	mean	min	M/M	mean	min	M/M	mean	min	M/M	mean	min	M/M
AVG [1]	0.177	0.057	0.182	0.198	0.054	0.147	**0.294**	0.188	0.470	**0.457**	0.288	0.471
FAI [1]	0.171	*0.084*	**0.321**	0.176	0.076	**0.259**	0.208	0.139	0.498	0.312	0.190	0.428
SPGreedy [6]	0.181	0.074	0.242	0.197	0.074	0.216	0.273	0.184	0.503	0.404	0.267	0.492
GreedyLM [7]	*0.187*	0.081	0.268	*0.205*	*0.079*	*0.222*	0.279	*0.204*	*0.569*	0.437	*0.315*	*0.565*
XPO [11]	0.201	0.087	*0.271*	0.218	0.082	*0.222*	0.250	0.161	0.475	0.390	0.235	0.433
GFAR [5]	**0.202**	**0.092**	0.287	**0.225**	**0.087**	0.232	0.247	0.162	0.489	0.392	0.236	0.433
FuzzDA [8]	0.178	0.068	0.230	0.199	0.067	0.191	0.293	0.220	0.604	0.455	0.336	0.591
EP-FuzzDA [8]	0.152	0.059	0.235	0.165	0.055	0.186	*0.280*	**0.251**	**0.821**	*0.438*	**0.369**	**0.730**

further modifications. Finally, for de-biasing evaluation scenario we considered the following γ hyperparameter (Eq. 3) values: $\gamma \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$. This slightly extends the range evaluated in the original study [12], so that we can get sufficient "overview" on how this parameter affects results.

Two evaluation metrics are considered: normalized discounted cumulative gain (nDCG) and average relevance score (AR), both evaluated on top-20 recommendations. For each list of recommendations L_G, we evaluate these metrics for all group members $u \in \mathcal{G}$ and collect three aggregated per-group statistics: *mean* per-user scores for relevance evaluation, *min*imal user's score (i.e., least misery fairness metric as used in [5, 11, 7]) and the ratio between minimal and maximal scores (as used in [5, 7]). In results, we report average values of these statistics for individual group types and group sizes. We often denote these metrics as a pair of base metric and per-group aggregation, e.g. (AR, mean).

3.3. Results

Table 1 depicts the results of coupled and decoupled evaluation scenarios for similar groups of size $s = 8$. We can observe that both types of evaluation give us highly different ranking of best

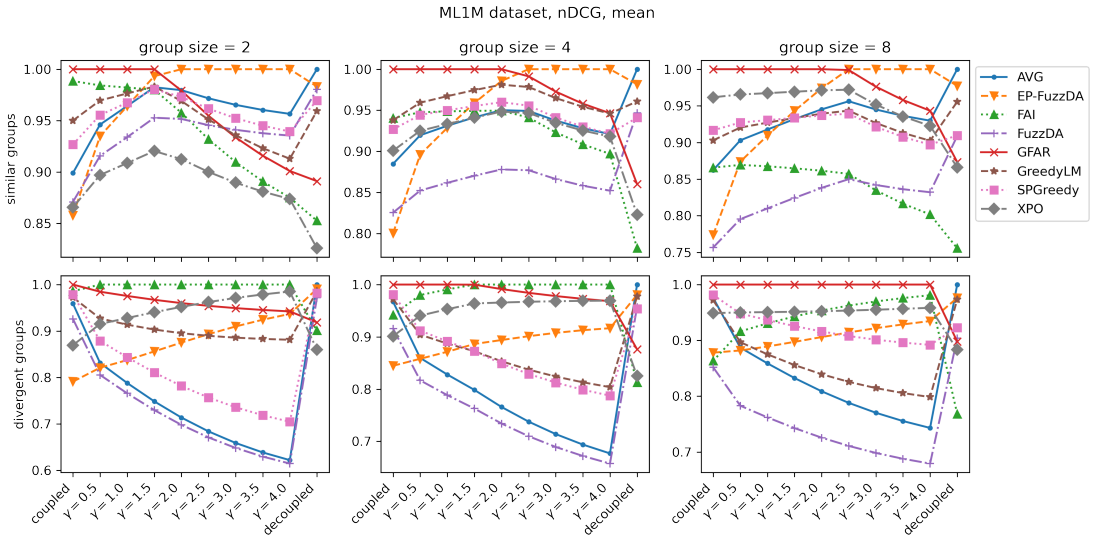

Figure 2: Relative results w.r.t. mean nDCG on ML1M dataset. All results are linearly normalized to the best-performing approach. $\gamma = x$ labels denote de-biasing evaluation with specific γ hyperparameters.

approaches. For *coupled* evaluation, GFAR and XPO compete for the best approach (seemingly, GFAR is slightly better). FAI algorithm provide good ratios of minimal vs. maximal per-group performance, but its mean per-group performance is inferior to the previous two examples. *Decoupled* evaluation provides a different view: EP-FuzzDA receives the best minimal per-group scores as well as minimal vs. maximal ratios followed by GreedyLM and AVG algorithms. For the mean scores, naturally, AVG is the best approach followed by EP-FuzzDA and GreedyLM (ML1M dataset), resp. FuzzDA and EP-FuzzDA (KGRec dataset). The difference between coupled and decoupled approach can be further illustrated by the Pearson's correlation, which is -0.32 and -0.29 for mean AR and mean nDCG on ML1M dataset respectively. Note that although we only depict the results for similar groups of size $s = 8$, results for other sizes of the *similar* groups exhibited analogical levels of contradiction. As for the divergent groups, the level of agreement between *coupled* and *decoupled* evaluations was higher in general (results were positively correlated in most cases), but the disagreement on the best-performing approaches prevails.

Figures 2 and 3 depict results of group RS aggregators w.r.t. coupled, decoupled and all variants of de-biasing scenarios. Because the scale of results w.r.t. individual evaluation scenarios differs greatly, we depict their normalized comparison. For the sake of space, we only show two of the evaluated metrics, additional figures are available from supplementary materials.

Results of similar groups exhibited a clear dependence between algorithm's tendency to recommend per-user best items or items with overall agreement and the γ hyperparameter of the de-biasing evaluation. The performance of algorithms preferring per-user best (FAI, GFAR, XPO) decreased with the increasing γ. On the other hand, algorithms preferring items with overall agreement (AVG, EP-FuzzDA, FuzzDA) gradually improved their relative performance with increasing γ values. The transition between the de-biasing evaluation with $\gamma = 4.0$ and

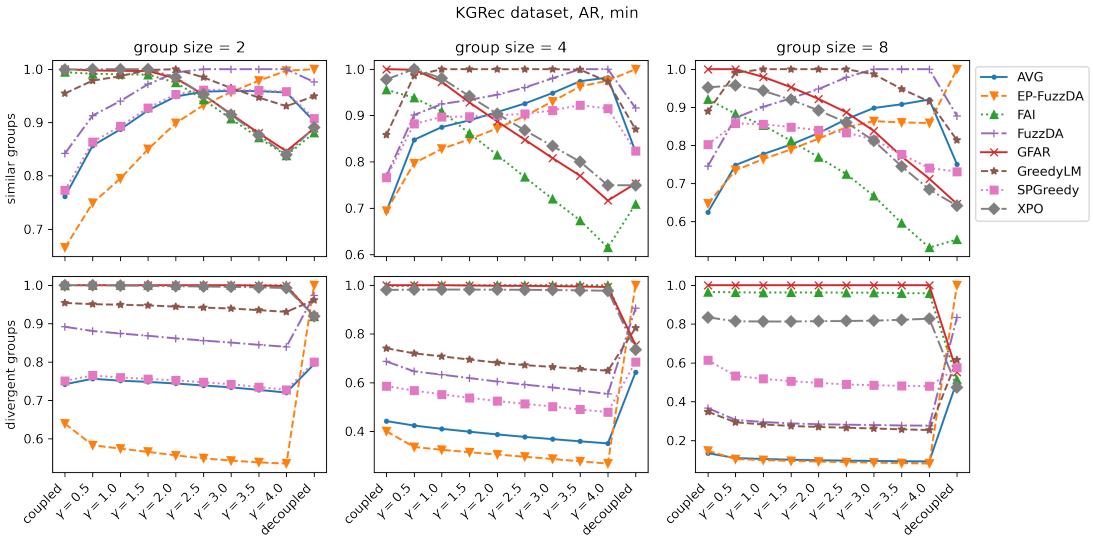

Figure 3: Relative results w.r.t. minimal per-group values of average relevance (AR) on KGRec dataset. All results are linearly normalized to the best-performing approach. $\gamma = x$ labels denote de-biasing evaluation with specific γ hyperparameters.

the decoupled evaluation was not as smooth as for the individual de-biasing scenarios, but remain fairly close (e.g., Kendall τ score was 0.71, 0.64 and 0.71 for KGRec dataset, (AR, min) metric and group sizes of 2, 4 and 8 respectively). Similar results were obtained for both datasets, and per-group mean and minimal values of both metrics.

Results w.r.t. divergent groups were less clear. In ML1M dataset and mean nDCG (Figure 2), we can observe some of the tendencies from similar groups. For instance, GFAR performance gradually decrease with increasing γ and the opposite applies for EP-FuzzDA. In contrast, the performance decrease of AVG, FuzzDA and SPGreedy was in contradiction to the results of corresponding similar groups. Also, there was almost complete shift of algorithm's relative ordering between de-biasing, $\gamma = 4.0$ and decoupled evaluation scenarios (Kendall τ scores of -0.43, -0.57 and -0.43 for group sizes $s = 2$, 4 and 8 respectively). Similar results were obtained also for divergent groups on ML1M dataset and (AR, mean) metric.

Results w.r.t. per-group minimal AR on KGRec dataset were rather indifferent of the choice of γ and highly resembles those of the coupled evaluation. Similarly as in the previous case, there was almost complete shift of algorithm's ordering between de-biasing, $\gamma = 4.0$ and decoupled evaluation scenarios. Similar results were obtained also for divergent groups on ML1M dataset and (AR, min), (AR, min/max) and (nDCG, min) metrics as well as KGRec dataset and (AR, mean), (AR, min/max) and (nDCG, min) metrics.

4. Discussion and Future Work

In this paper, we compared several off-line evaluation protocols for group RS aggregation strategies. First, we showed that two widely adopted strategies, *coupled* evaluation and *decoupled*

evaluation can lead to highly contradictory results. Then, we utilized de-biasing evaluation with the aim to provide a smooth transition between coupled and decoupled evaluation strategies. De-biasing strategy is coupled in nature, but it introduces an inverse propensity based normalization of results that mitigate one of the major flaws in coupled evaluation: assumption on randomly missing feedback.

This task was partially achieved as long as groups of similar users are considered. We observed increase of performance w.r.t. γ values for strategies that prefer items with overall agreement rather than per-user bests (AVG, FuzzDA, EP-FuzzDA) and in contrast, performance of strategies that tends to propose per-user best items (FAI, GFAR, XPO) drops with increasing γ. These results were mostly consistent with the ones given by decoupled evaluation strategy.

We believe that the main factor behind the observed behavior is the combination of MNAR dataset and popularity-biased underlying RS, which boosted the chance that top recommended items were observed (above the relative difference in estimated preferences) and which is gradually penalized by the de-biasing evaluation scenario. This gives a chance to recommend less popular, but more agreeable items.

For divergent groups of users, one possible explanation is that commonly preferred items were simply not present in the test sets.[4] Therefore, penalizing recommendations of too popular items as in de-biasing scenario would not change the results much, because there are simply no good alternatives instead of per-user best items.[5] Decoupled evaluation, in theory, can bridge this problem as the underlying RS can discover mutually agreeable items that are not present in the historical data. This may be the cause of the large performance jump between the de-biasing $\gamma = 4.0$ and decoupled evaluation scenarios for divergent groups. However, the same effect could be also observed if the underlying RS overestimates the true preferences of users. ALS MF utilized in this work rarely ever provided negative $\hat{r}_{u,i}$ ratings and as no ratings' post-processing was performed, most of the items were considered as (at least) mildly preferred by users. Many RS would behave in a similar fashion. Therefore, detecting and counter-measuring the overestimation bias is one direction of our future work.

To conclude, the main message of this paper is that performance of group RS strategies can highly depend on the considered evaluation scenario. So, before applying any particular evaluation strategy, authors should carefully consider intended goals of the proposed algorithm and its compliance with the possible biases introduced by the evaluation process.

We consider our work as rather preliminary and there are numerous possible extensions. First, we only evaluated results w.r.t. one underlying RS. We plan to experiment with other recommenders to corroborate our findings. Especially, it would be interesting to observe the results w.r.t. some content-based RS that should be less prone to the popularity bias and e.g. variants of item KNN to decrease the overestimation bias. Also, instead of applying de-biasing evaluation strategy, it is possible to bridge the results of coupled and decoupled evaluation by utilizing only a portion of RS prediction (e.g. only those with the highest estimated relevance). This could be another option to reduce the effect of overestimation bias.

Our long-term future work focus on real-world performance of group recommenders. For

[4] As the groups were assembled w.r.t. the level of disagreement among users, it is not an unrealistic assumption. Nonetheless, further validation is needed.

[5] As long as some popularity de-biasing strategy, e.g. [15] is not considered, which was not the case of evaluated approaches.

instance, one question risen by this study is whether the real-world user groups are intrinsically rather similar or diverse. This knowledge can greatly affect the success of agreement oriented vs. per-user-best oriented designs of group RS aggregators. Also the relation between coupled/decoupled/de-biasing evaluations and on-line performance is currently unknown. Therefore, another direction of our future work is to conduct a realistic user study that could both provide some insights on user group formations as well as evaluate the on-line performance of group recommenders.

Acknowledgments

The work on this paper has been supported by Czech Science Foundation project GACR-19-22071Y and by Charles University grant SVV-260588. Additional results can be obtained from https://github.com/lpeska/Perspectives-RecSys2021.

References

[1] J. Masthoff, Group Recommender Systems: Combining Individual Models, Springer US, Boston, MA, 2011, pp. 677–702. URL: https://doi.org/10.1007/978-0-387-85820-3_21. doi:10.1007/978-0-387-85820-3_21.

[2] A. Crossen, J. Budzik, K. J. Hammond, Flytrap: Intelligent group music recommendation, in: Proceedings of the 7th International Conference on Intelligent User Interfaces, IUI '02, Association for Computing Machinery, New York, NY, USA, 2002, p. 184–185. URL: https://doi.org/10.1145/502716.502748. doi:10.1145/502716.502748.

[3] J. F. McCarthy, Pocket restaurantfinder: A situated recommender system for groups, in: Workshop on Mobile Ad-Hoc Communication at the 2002 ACM Conference on Human Factors in Computer Systems, 2002.

[4] L. Quijano-Sánchez, B. Díaz-Agudo, J. A. Recio-García, Development of a group recommender application in a social network, Knowledge-Based Systems 71 (2014) 72–85. URL: https://www.sciencedirect.com/science/article/pii/S095070511400197X. doi:https://doi.org/10.1016/j.knosys.2014.05.013.

[5] M. Kaya, D. Bridge, N. Tintarev, Ensuring fairness in group recommendations by rank-sensitive balancing of relevance, in: Fourteenth ACM Conference on Recommender Systems, RecSys '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 101–110. URL: https://doi.org/10.1145/3383313.3412232. doi:10.1145/3383313.3412232.

[6] D. Serbos, S. Qi, N. Mamoulis, E. Pitoura, P. Tsaparas, Fairness in package-to-group recommendations, in: Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 371–379. URL: https://doi.org/10.1145/3038912.3052612. doi:10.1145/3038912.3052612.

[7] L. Xiao, Z. Min, Z. Yongfeng, G. Zhaoquan, L. Yiqun, M. Shaoping, Fairness-aware group recommendation with pareto-efficiency, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys '17, Association for Computing Machinery, New York,

NY, USA, 2017, p. 107–115. URL: https://doi.org/10.1145/3109859.3109887. doi:10.1145/3109859.3109887.

[8] L. Malecek, L. Peska, Fairness-preserving group recommendations with user weighting, in: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 4–9. URL: https://doi.org/10.1145/3450614.3461679. doi:10.1145/3450614.3461679.

[9] L. Quijano-Sanchez, J. A. Recio-Garcia, B. Diaz-Agudo, G. Jimenez-Diaz, Social factors in group recommender systems, ACM Trans. Intell. Syst. Technol. 4 (2013). URL: https://doi.org/10.1145/2414425.2414433. doi:10.1145/2414425.2414433.

[10] L. Quijano-Sánchez, J. A. Recio-García, B. Díaz-Agudo, Modelling hierarchical relationships in group recommender systems, in: E. Hüllermeier, M. Minor (Eds.), Case-Based Reasoning Research and Development, Springer International Publishing, Cham, 2015, pp. 320–335.

[11] D. Sacharidis, Top-n group recommendations with fairness, in: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1663–1670. URL: https://doi.org/10.1145/3297280.3297442. doi:10.1145/3297280.3297442.

[12] L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie, D. Estrin, Unbiased offline recommender evaluation for missing-not-at-random implicit feedback, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 279–287. URL: https://doi.org/10.1145/3240323.3240355. doi:10.1145/3240323.3240355.

[13] I. Pilászy, D. Zibriczky, D. Tikk, Fast als-based matrix factorization for explicit and implicit feedback datasets, in: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 71–78. URL: https://doi.org/10.1145/1864708.1864726. doi:10.1145/1864708.1864726.

[14] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 10th International Conference on World Wide Web, WWW '01, Association for Computing Machinery, New York, NY, USA, 2001, p. 285–295. URL: https://doi.org/10.1145/371920.372071. doi:10.1145/371920.372071.

[15] H. Abdollahpouri, R. Burke, B. Mobasher, Controlling popularity bias in learning-to-rank recommendation, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 42–46. URL: https://doi.org/10.1145/3109859.3109912. doi:10.1145/3109859.3109912.

[16] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2015).

[17] S. Oramas, V. C. Ostuni, T. D. Noia, X. Serra, E. D. Sciascio, Sound and music recommendation with knowledge graphs, ACM Trans. Intell. Syst. Technol. 8 (2016). URL: https://doi.org/10.1145/2926718. doi:10.1145/2926718.