# The Effect of Similarity Metric and Group Size on Outlier Selection & Satisfaction in Group Recommender Systems

Patrik Dokoupil
patrik.dokoupil@matfyz.cuni.cz
Faculty of Mathematics and Physics, Charles University, Prague
Czechia

Ladislav Peska
ladislav.peska@matfyz.cuni.cz
Faculty of Mathematics and Physics, Charles University, Prague
Czechia

## ABSTRACT

Group recommender systems (GRS) are a specific case of recommender systems (RS), where recommendations are constructed to a group of users rather than an individual. GRS has diverse application areas including trip planning, recommending movies to watch together, or music in shared environments. However, due to the lack of large datasets with group decision-making feedback information, or even the group definitions, GRS approaches are often evaluated offline w.r.t. individual user feedback and artificially generated groups. These synthetic groups are usually constructed w.r.t. pre-defined group size and inter-user similarity metric. While numerous variants of synthetic group generation procedures were utilized so far, its impact on the evaluation results was not sufficiently discussed. In this paper, we address this research gap by investigating the impact of various synthetic group generation procedures, namely the usage of different user similarity metrics and the effect of group sizes. We consider them in the context of "outlier vs. majority" groups, where a group of similar users is extended with one or more diverse ones. Experimental results indicate a strong impact of the selected similarity metric on both the typical characteristics of selected outliers as well as the performance of individual GRS algorithms. Moreover, we show that certain algorithms better adapt to larger groups than others.

## CCS CONCEPTS

• **Information systems** → **Recommender systems**.

## KEYWORDS

Group Recommender systems; Synthetic groups construction; User similarity metrics

**ACM Reference Format:**
Patrik Dokoupil and Ladislav Peska. 2023. The Effect of Similarity Metric and Group Size on Outlier Selection & Satisfaction in Group Recommender Systems. In *UMAP '23 Adjunct: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23 Adjunct), June 26–29, 2023, Limassol, Cyprus.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3563359.3597386



*UMAP '23 Adjunct, June 26–29, 2023, Limassol, Cyprus*

ACM ISBN 978-1-4503-9891-6/23/06.
https://doi.org/10.1145/3563359.3597386

## 1 INTRODUCTION

Group recommender systems (GRS) are a sub-domain of recommender systems (RS) that focus on serving recommendations to groups of users instead of individual users. Although the widespread of GRS in production environments is presently limited, there are numerous promising application areas where activities (and thus, recommendations) are typically conducted in groups. Some examples are traveling, dining, attending various cultural events, or being exposed to some background music in shared environments. Nevertheless, the focus on groups instead of individuals adds one more layer of complexity to the recommendation process. Although there are some proposed end-to-end solutions [16, 24], typical GRS approaches are applied as post-processing on top of the preferences of individual group members. One of the common requirements on GRS functionality is to maintain some degree of fairness among the group members, i.e., to ensure that all group members are reasonably satisfied with the provided recommendations.

What makes this task more challenging (both conceptually and in terms of implementation) is the heterogeneity of the group's composition. This could come in many different flavors—the group may consist of divergent users having opposite preferences [12], there may be several homogeneous subgroups that are different from each other [20], various dominance or trust relations may affect group dynamics [4, 5, 21] and so on. In such an environment, it is not only difficult to provide reasonable algorithmic solutions, but even to properly define the target metric(s) to optimize.

As an example, consider a group of 10 users whom we aim to recommend a list of 10 items. Nine group members are highly similar, and the same set of recommendations can satisfy them to an equal degree. However, the last group member is an outlier whose preferences diverge from the rest of the group, and none of the previously mentioned items are satisfactory. Should the fair system aim to a) satisfy each user to the same degree, i.e., dedicate half of the recommending slots to the outlier user, b) dedicate a proportional fraction of the recommending slots to the outlier user, or c) discard his/her preferences for the sake of the majority. While the resolution of this conceptual problem is out of scope of this paper, it may still be interesting to highlight what concept of fairness is closest to individual GRS algorithms.

Another obstacle lies in the lack of sufficiently large datasets that would describe the groups' decision processes, or at least the groups' composition. Even though there are some datasets containing group decisions [6, 7], they are of limited size and, consequently, do not offer much variability in the groups' composition. To overcome this shortage, researchers often resort to the usage of datasets with

individual user preferences[1], on top of which the synthetic groups are generated. A typical groups generation procedure is based on some notion of user-user similarity and constructs a set of groups of some prescribed properties. Some examples of such prescriptions are: generate groups of mutually similar [12, 13], mutually divergent [12, 23] or random [12] users of certain sizes, groups with two or more divergent subgroups [20] etc. As for the similarity metric, different approaches utilized, e.g., users' rating vectors [12, 17], or more complex interaction graphs [20], their trained embeddings [13], or their content-based features [15].

Note that the usability of synthetic groups stands and falls with the ability of the generating procedure to provide an unbiased sample of the (latent) distribution of real-world groups for the GRS use-case at hand. Nonetheless, as the distributions of real-world groups are unknown for the vast majority of GRS tasks, different group generation procedures are mostly utilized ad hoc, without much discussion on their appropriateness. We believe this is a serious threat affecting the credibility of reported results throughout the GRS research domain.

Unfortunately, to the best of our knowledge, the source data needed to arbitrate the appropriateness of individual group generation procedures is currently missing, and we are not aware of any study aiming to analyze this phenomenon. Instead, in this paper, we merely aim on evaluating the level of impact that the groups generation procedures have on the groups' properties and the results of different GRS techniques. In particular, we focus on groups with one majority subgroup and one or more outlier members. For such a prescription, we test various types of user-user similarity metrics as well as various sizes of groups. In the subsequent analysis, we aim to answer the following research questions:

- **RQ1:** *Does the choice of similarity metric affects some properties of generated groups, or their respective majority and outlier sub-groups?*
- **RQ2:** *Does the particular choice of a similarity metric or group size affect the performance of GRS algorithms? Does this change if considered from the majority or outlier points of view?*

## 2 RELATED WORK

A lot of research has been done in terms of designing new GRS algorithms. The early GRS algorithm proposals focused on the per-item aggregation scores (i.e., not considering the context of other recommended items). Some of the well-known examples are *Least Misery*, *Borda Count*, *Average*, *Multiplicative* etc. [18]. More recently, GRS-related research shifted towards the problem of list-level algorithmic fairness w.r.t. individual group members (i.e., considering the impact of the whole list of recommendations, not just single items). Let us mention some examples. *GreedyLM* [23] is a greedy algorithm that attempts to achieve fairness by iteratively selecting items that maximize the linear combination of overall group satisfaction and the satisfaction of the least satisfied group member (w.r.t. the so-far constructed list). *GFAR* [12] algorithm aims to directly optimize a rank-sensitive fairness objective (i.e., maximizing the probability of all users to find at least one relevant hit in the list of recommendations). *EPFuzzDA* [17] aims to achieve a results-level proportionality w.r.t. weights assigned to each user via a modified mandates allocation algorithm.

There are several existing works that performed evaluations on synthetically generated groups. In [12], authors considered similar, divergent, and random groups of sizes between 2 and 8. Inspired by [2, 10], they use the Pearson correlation coefficient to measure similarity between the user's ratings and build groups incrementally by seeking users whose similarity is above/below a certain threshold for similar/divergent groups. In random groups, all users have a uniform probability of being selected. The same approach was also used in [3]. In [23], similar, diverse, and random groups were also constructed, but cosine similarity was utilized, and missing ratings were substituted by its estimation w.r.t. BPR MF [22]. In contrast, [13, 14] utilize similarity metrics based on content-based embedding vectors and user demographics, respectively. However, a discussion on the appropriateness of the group generation process, or in particular the utilized similarity metrics, is sketchy or missing completely in the related papers.

## 3 SYNTHETIC GROUP CONSTRUCTION

Let us now describe the considered group generation procedures. We largely base the protocol on [3], with a few extensions. In particular, 200 similar groups were generated for each of the group sizes $|G| \in \{4, 8, 32, 100\}$. The procedure was incremental, starting with $G = \{u\}$ for some random $u$. Then, a partially generated group $G$ was extended with a random user $u \in U_c$, where $U_c$ is a set of all users whose mean similarity towards members of $G$ is above a certain threshold: $U_c \stackrel{\text{def}}{=} \{u \in \mathcal{U} \mid \frac{\sum_{u_g \in G} sim(u,u_g)}{|G|} \geq tresh_{sim}\}$. The $sim$ is the actual inter-user similarity metric (see the list below), and $thresh_{sim}$ is the similarity threshold. For each $sim$ metric, we set the threshold to $99th - percentile$ of values in the respective similarity matrices. This should ensure that only really similar users are selected, irrespective of ranges of different $sim$ metrics.

We further alter generated *similar* groups to also contain an *outlier* sub-group. In particular, we removed the first $n_{G,outliers} \stackrel{\text{def}}{=} \lceil |G|/10 \rceil$ members of each similarity group and replaced them with different users as follows. The first replacement, $o_1$, was selected as the user with minimal mean similarity towards the remaining group members. The remaining outliers were selected as users with the highest similarity towards $o_1$. This effectively constructs two sub-groups within each group: *majority* and *outliers*.

The similarity metrics that were used come from two categories—*data-based* metrics calculated on the raw ratings data, and *feature-based* metrics calculated on the learned user embeddings of the underlying RS. Overall, the following variants were considered:

- *data-based*: Pearson correlation coefficient (PCC) among users, calculated either on full or training part of interaction matrix (denoted as *PCC_full* and *PCC_train* respectively).
- *feature-based*: L2 and Cosine similarity calculated on the learned user features from either Biased Matrix Factorization [25] (*L2_biasMF*, and *COS_biasMF*) or Implicit Matrix Factorization [11] (*L2_implMF*, and *COS_implMF*). The L2-based similarities were calculated by transforming the L2 distance as $\frac{1}{1+distance_{L2}}$.

[1] Usually based on some individual recommendation scenario, such as MovieLens [9].

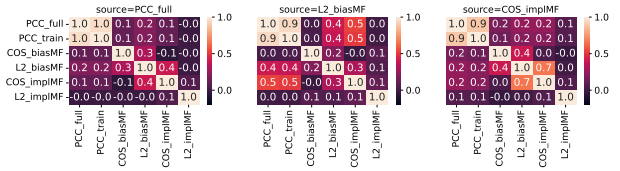


**Figure 1: The PCC matrices for $sim(outlier, majority)$ for groups constructed via *PCC_full*, *L2_biasMF*, and *COS_implMF*.**

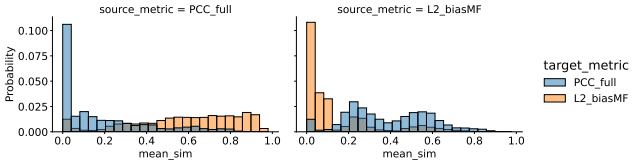


**Figure 2: Histogram of similarities for pairs of outlier-majority sub-group members. Results for *PCC_full* and *L2_biasMF* are depicted.**

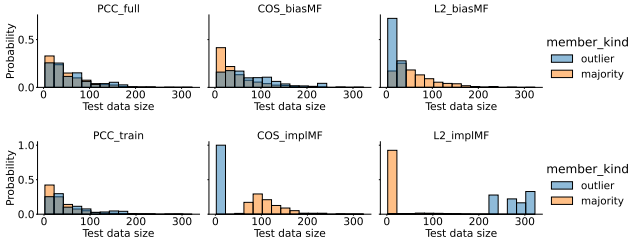


**Figure 3: Histogram of Test set sizes for the members of outlier and majority sub-groups. Results are separated w.r.t. source metrics.**

## 4 EVALUATION AND RESULTS

Similarly to the groups generation procedure, the evaluation protocol was largely based on [3]. In particular, we utilized MovieLens1M dataset [9][2], coupled evaluation approach [19] with 80:20 stratified train-test split, and Biased MF [25] as a source of user's individual preferences.[3] Subsequently, for each group, we generated a list $L$ of top-10 recommendations using the following GRS algorithms: Multiplicative (*MUL*), Additive (*ADD*), Least misery (*LMS*), Borda count (*BDC*), and Most pleasure (*MPL*) [18] as representatives of item-wise aggregators and Fairness (*FAI*, [18]), *EPFuzzDA* [17], GFAR [12], and *GreedyLM* [23] as representatives of list-wise approaches. We considered a simple proxy for user satisfaction: $precision@\text{top-}10 \stackrel{\text{def}}{=} \frac{|\{i,\, i \in Test_u \cap L\}|}{|L|}$, where $Test_u$ represent test set items with known feedback from user $u$ and $L$ is the list of top-10 recommended items for the group.

### 4.1 Results

Starting with **RQ1**, we evaluated to what extent individual similarity metrics correspond to each other, while assessing group members' similarity. In particular, we considered groups constructed via different similarity metrics[4] and then re-evaluate the similarity of all outlier-majority pairs w.r.t. all similarity metrics. Figure 1 depicts Pearson's correlations between similarity estimations of all pairs of metrics. Notably, when groups are generated w.r.t. *PCC_full*, only *PCC_train* produces a highly correlated view on the group member's similarity. Similarly, for groups based on *COS_implMF*, only *L2_biasMF* produced a correlated view, while for groups based on *L2_biasMF*, none of the other metrics is sufficiently correlated with the original one. Overall, *PCC_full* and *PCC_train* exhibited highly similar results throughout all group definitions. Apart from that, only moderate correlations were observed, which further differs based on the groups' generation procedure. Notably, many pairs of similarity metrics produced results with correlations close to zero. This indicates a severe problem, i.e., while using one perspective to define similar or diverse users, this does not hold if another perspective is applied.

We dug a bit deeper into the topic and analyzed the histograms of mean similarities of outliers towards majority group members. Figure 2 depicts variants, where *PCC_full* and *L2_biasMF* serve both as *source metrics* (i.e., the metric based on which the groups were generated) and *target metrics* (i.e., metrics whose histograms are depicted)[5]. Note that in order to ensure inter-metric comparability, similarity matrices were normalized using the empirical cumulative distribution function. An interesting observation is that for groups generated w.r.t. *PCC_full* metric (the most common approach in the literature), outliers are actually often considered as highly similar to the majority members w.r.t. *L2_biasMF*[6]. Combined with the fact that *BiasMF* is often used as a source of single-user relevance estimations in the downstream GRS, this threatens the integrity of GRS performance evaluation. The fact that outliers have highly similar embedding vectors to the majority group members implies a highly similar list of estimated preferences (i.e., there is no distinction between outlier and majority users from the RS's perspective). This makes the group recommendation task much simpler than intended, and unrealistically positive results may be expected from a decoupled evaluation scenario [19], but also from a coupled one if a popularity bias [1] is present.

Finally, we focused on whether the outliers vs. majority members differ in some of their general features. Indeed, we observed major discrepancies in terms of the rating profile sizes for majority vs. outlier sub-groups (see Figure 3). For instance, *COS_implMF* constructs groups where outliers have extremely small test sets as compared to the majority sub-group. A similar discrepancy was also observed for *L2_biasMF*, although not as extreme. On the other hand, *L2_implMF* often assigns users with extremely small test sets during majority group construction, while for outliers, users

[2]With the same pre-processing as used in [3].
[3]We used Lenskit [8] implementation of Biased MF and trained the model with a feature size = 30, regularization = 0.1, and 20 iterations.
[4]For the sake of space, we only report on *PCC_full*, *L2_biasMF*, and *COS_implMF* groups. Nonetheless, other results were analogical.
[5]Other combinations of source and target metrics are available from supplementary materials.
[6]Similar results were also obtained for other feature-based metrics.


**Figure 4:** *Precision*@10 **for outlier and majority point of view, for various configurations. Rows correspond to similarity metrics and columns to different group sizes** (4, 8, 32, 100)**. EPFuzzDA is denoted as EPFDA and GreedyLM as GLM due to space constraints.**

with very large profiles are often selected. Obviously, having sub-groups with differently-sized user profiles may affect their relative comparisons, because it is much easier to satisfy users with larger profiles (considering metrics such as *precision*@top-k). From this perspective, *PCC_full*, *PCC_train*, and also *COS_biasMF* derive more suitable distributions of outlier and majority group members.

Focusing on **RQ2**, we assessed the performance of GRS algorithms (w.r.t. *precision*@top-10) from the perspective of both outlier and majority sub-groups. Figure 4 depicts the comparison between the mean performance of outlier and majority sub-groups as well as overall results.

Notably, for groups generated w.r.t. *L2_implMF*, and partially also *PCC_full*, *PCC_train*, and *COS_biasMF*, many GRS variants exhibited higher performance on outliers than on majority sub-group. On the other end of the spectrum, for *COS_implMF* the performance of all GRS is close to zero for outlier sub-groups. As for *L2_implMF* and *COS_implMF*, this is clearly an artifact of the test set sizes. We believe the cause is the same also for the other cases. Despite the differences being much more subtle, the mean test set size for majority sub-group members was smaller than those of the outliers sub-group. For *L2_implMF* and *COS_biasMF* groups, we can highlight the performance of *MPL*, *GFAR* and *FAI* algorithms, which were able to provide satisfactory results (to some extent) also for the disadvantaged sub-groups (i.e., outliers for *COS_implMF* and majority for *L2_implMF*). Overall, we can see a clear impact of the test set sizes on the performance across all evaluated GRS, similarity metrics, and group sizes. We would like to dedicate one direction of our future work to provide a more robust analysis of this phenomenon and to propose appropriate measures to counter this bias in GRS evaluation.

In terms of group size, the general trend is that the performance decreases with increasing group sizes. This is not surprising, but we also found several exceptions, e.g. the performance of FAI algorithm on *PCC_full*-based groups. Overall, *MPL*, *FAI*, and *GFAR* algorithms exhibited the highest stability of their performance w.r.t. both sub-groups across different group sizes. Considering, e.g., *PCC_full*, the

performance of the aforementioned algorithms is on par or inferior to other variants for $|G| = 4$. However, while it remains fairly stable also for $|G| \in \{8, 32, 100\}$, the performance of other algorithms drops significantly. This holds for both outlier and majority subgroups. These observations may indicate a potential concern with GRS evaluation. Relying solely on small group sizes for evaluation may result in biased outcomes, where certain algorithms are deemed superior or equivalent to others without adequately considering their scalability issues.

We also focused on the discrepancies in the outlier-wise and overall performance of individual GRS. While the results corresponded in most cases, there were some notable exceptions. Let us mention e.g. *L2_biasMF* with $|G| = 8$, where the overall performance of *MPL* and *GFAR* was similar to *FAI*, but *FAI* achieved significantly better performance on outliers. A similar effect can be seen when comparing *MPL* with *GFAR* for *COS_implMF* and $|G| \in \{32, 100\}$. In contrast, despite the overall performance of FAI was highest for *PCC_full* and *PCC_train* on $|G| = 4$, it was considerably outperformed outlier-wise by several GRS. This corresponds with the fact that FAI maps each user to an equal volume of recommendation slots, rather than aiming to provide an equal outcome for all users (as most of the other GRS). Nonetheless, we did not observe sufficiently stable trends w.r.t. outlier vs. overall performance to make some definite conclusions on this matter.

## 5 CONCLUSION

In general, our findings indicate that the GRS evaluation process may be substantially affected by the synthetic groups' generation procedure, namely by the choice of the similarity metric and considered group sizes. We found that individual similarity metrics mostly do not correlate with each other and may produce groups with highly different statistics (e.g., varying profile sizes of group members). This may, in turn, impact the performance of individual GRS but also the relative comparison of considered sub-groups (e.g., outlier and majority sub-groups in our case). Overall, by this work, we would like to stress the importance of verifying the properties of generated synthetic groups (and their respective sub-groups) and highlight possible impacts this may have on the off-line GRS evaluation.

### 5.1 Limitations and future work

This paper is subject to several limitations. Firstly, the experimental design only encompasses a single type of outlier group and employs a simplistic formula to determine the number of outliers per group. Additionally, a fixed group construction approach was utilized. Secondly, although six similarity metrics were compared, there are numerous other options that require exploration and assessment. It should be noted that the aim of this paper was not to identify the best similarity metric but rather to highlight the potential consequences of choosing some specific metric. We plan to address these limitations in our future work and expand current experiments on additional datasets, using additional evaluation protocols and metrics and also incorporating various debiasing techniques.

## ACKNOWLEDGMENTS

This paper has been supported by Charles University grant SVV-260698, and by Charles University Grant Agency (GA UK) project number 188322. Supplementary materials are available from https://github.com/pdokoupil/gmap2023.

## REFERENCES

[1] Abdul Basit Ahanger, Syed Wajid Aalam, Muzafar Rasool Bhat, and Assif Assad. 2022. Popularity Bias in Recommender Systems - A Review. In *Emerging Technologies in Computer Engineering: Cognitive Computing and Intelligent IoT*, Valentina E. Balas, G. R. Sinha, Basant Agarwal, Tarun Kumar Sharma, Pankaj Dadheech, and Mehul Mahrishi (Eds.). Springer International Publishing, Cham, 431–444.

[2] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. 2010. Group Recommendations with Rank Aggregation and Collaborative Filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 119–126. https://doi.org/10.1145/1864708.1864733

[3] Francesco Barile, Amra Delic, and Ladislav Peska. 2022. Tutorial on Offline Evaluation for Group Recommender Systems. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 702–705. https://doi.org/10.1145/3523227.3547371

[4] David Contreras, Maria Salamó, and Ludovico Boratto. 2021. Integrating Collaboration and Leadership in Conversational Group Recommender Systems. *ACM Trans. Inf. Syst.* 39, 4, Article 41 (aug 2021), 32 pages. https://doi.org/10.1145/3462759

[5] Amra Delic, Judith Masthoff, Julia Neidhardt, and Hannes Werthner. 2018. How to Use Social Relationships in Group Recommenders: Empirical Evidence. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, New York, NY, USA, 121–129. https://doi.org/10.1145/3209219.3209226

[6] Amra Delic, Julia Neidhardt, Thuy-Ngoc Nguyen, and Francesco Ricci. 2018. An observational user study for group recommender systems in the tourism domain. *Information Technology & Tourism* 19 (06 2018). https://doi.org/10.1007/s40558-018-0106-y

[7] Amra Delic, Julia Neidhardt, Thuy Ngoc Nguyen, Francesco Ricci, Laurens Rook, Hannes Werthner, and Markus Zanker. 2016. Observing Group Decision Making Processes. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 147–150. https://doi.org/10.1145/2959100.2959168

[8] Michael D. Ekstrand. 2020. LensKit for Python: Next-Generation Software for Recommender Systems Experiments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2999–3006. https://doi.org/10.1145/3340531.3412778

[9] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[10] Daniel Herzog and Wolfgang Wörndl. 2019. A User Study on Groups Interacting with Tourist Trip Recommender Systems in Public Spaces. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 130–138. https://doi.org/10.1145/3320435.3320449

[11] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*. 263–272. https://doi.org/10.1109/ICDM.2008.22

[12] Mesut Kaya, Derek Bridge, and Nava Tintarev. 2020. Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/3383313.3412232

[13] Ondrej Kaššák, Michal Kompan, and Mária Bieliková. 2016. Personalized hybrid recommendation for group of users: Top-N multimedia recommender. *Information Processing & Management* 52, 3 (2016), 459–477. https://doi.org/10.1016/j.ipm.2015.10.001

[14] Jae Kyeong Kim, Hyea Kyeong Kim, Hee Young Oh, and Young U. Ryu. 2010. A group recommendation system for online communities. *International Journal of Information Management* 30, 3 (2010), 212–219. https://doi.org/10.1016/j.ijinfomgt.2009.09.006

[15] Chintoo Kumar and C. Ravindranath Chowdary. 2021. Auto-detecting groups based on textual similarity for group recommendations. *CoRR* abs/2107.07284 (2021). arXiv:2107.07284 https://arxiv.org/abs/2107.07284

[16] Youfang Leng, Li Yu, and Xi Niu. 2022. Dynamically aggregating individuals' social influence and interest evolution for group recommendations. *Information*

*Sciences* 614 (2022), 223–239. https://doi.org/10.1016/j.ins.2022.09.058

[17] Ladislav Malecek and Ladislav Peska. 2021. Fairness-Preserving Group Recommendations With User Weighting. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 4–9. https://doi.org/10.1145/3450614.3461679

[18] Judith Masthoff and Amra Delić. 2022. *Group Recommender Systems: Beyond Preference Aggregation*. Springer US, New York, NY, 381–420. https://doi.org/10.1007/978-1-0716-2197-4_10

[19] Ladislav Peska and Ladislav Malecek. 2021. Coupled or Decoupled Evaluation for Group Recommendation Methods?. In *Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop 2021 co-located with the 15th ACM Conference on Recommender Systems (RecSys 2021), Amsterdam, The Netherlands, September 25, 2021 (CEUR Workshop Proceedings, Vol. 2955)*, Eva Zangerle, Christine Bauer, and Alan Said (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-2955/paper1.pdf

[20] Dong Qin, Xiangmin Zhou, Lei Chen, Guangyan Huang, and Yanchun Zhang. 2020. Dynamic Connection-Based Social Group Recommendation. *IEEE Transactions on Knowledge and Data Engineering* 32, 3 (2020), 453–467. https://doi.org/10.1109/TKDE.2018.2879658

[21] Lara Quijano-Sánchez, Juan A. Recio-García, and Belen Díaz-Agudo. 2015. Modelling Hierarchical Relationships in Group Recommender Systems. In *Case-Based Reasoning Research and Development*, Eyke Hüllermeier and Mirjam Minor (Eds.). Springer International Publishing, Cham, 320–335.

[22] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09)*. AUAI Press, Arlington, Virginia, USA, 452–461.

[23] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-Aware Group Recommendation with Pareto-Efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 107–115. https://doi.org/10.1145/3109859.3109887

[24] Song Zhang, Nan Zheng, and Danli Wang. 2022. GBERT: Pre-Training User Representations for Ephemeral Group Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 2631–2639. https://doi.org/10.1145/3511808.3557330

[25] Yunhong Zhou, Dennis M. Wilkinson, Robert Schreiber, and Rong Pan. 2008. Large-Scale Parallel Collaborative Filtering for the Netflix Prize. In *Algorithmic Aspects in Information and Management, 4th International Conference, AAIM 2008, Shanghai, China, June 23-25, 2008. Proceedings (Lecture Notes in Computer Science, Vol. 5034)*, Rudolf Fleischer and Jinhui Xu (Eds.). Springer, 337–348. https://doi.org/10.1007/978-3-540-68880-8_32