Algorithmically Mediated User Relations: Exploring Data's Relationality in Recommender Systems

Athina Kyriakou[1]* **Oana Inel**[1] **Asia J. Biega**[2] **Abraham Bernstein**[1]
[1]University of Zurich, Switzerland [2]Max Planck Institute for Security and Privacy, Germany
{kyriakou,inel,bernstein}@ifi.uzh.ch, asia.biega@mpi-sp.org

Abstract

Personalization services, such as recommender systems, operate on vast amounts of user-item interactions to provide personalized content. To do so, they identify patterns in the available interactions and group users based on pre-existing offline or online social relations, or algorithmically determined similarities and differences. We refer to the relations created between users based on algorithmically determined constructs as *algorithmically mediated user relations*. However, prior works in the fields of law, technology policy, and philosophy have identified a lack of existing algorithmic governance frameworks to account for this relational aspect of data analysis. Algorithmically mediated user relations have also not been adequately acknowledged in technical approaches, such as for data importance and privacy, where users are usually considered independent from one another. In this paper, we highlight this conceptual discrepancy in the context of recommendation algorithms and provide empirical evidence of the limitations of the user independence assumption. We discuss related implications and future practical directions for accounting for algorithmically mediated user relations.

1 Introduction

Personalization services collect increasing amounts of users' interactions in exchange for providing individualized content. Such services' underlying principle is that a user's previous preferences could be indicative of their future interests. Given that user interaction data are often sparse, personalization algorithms, such as collaborative filtering (CF) recommenders, seek to derive individuals' missing information based on data disclosed by others Aggarwal et al. (2016); Koren et al. (2021).

Prior literature has described ways in which data collected from an individual could be used to derive missing information about others Kammourieh et al. (2017); Barocas and Levy (2020). Firstly, undisclosed interactions can be inferred based on users' offline or online social relations Humbert et al. (2019); Parsons and Viljoen (2023), given that these relations could influence users' behavior Kempe et al. (2003); Li et al. (2023). Secondly, in the absence of a social network, a user's missing interaction could be drawn from others who are considered similar to them and for whom the interaction is known. This results in a data dependency between the target user and those considered similar to them, which depends on how user similarity and difference are defined by the algorithm used for data analysis and inference Viljoen (2021); Wachter (2022). We refer to these data dependencies created between users due to algorithmic processing as *algorithmically mediated user relations*.

Understanding the nature of algorithmically mediated user relations would contribute to addressing individual- and population-level harms related to privacy Barocas and Levy (2020), personalization Gordon-Tapiero et al. (2022), and discrimination Wachter (2022). This is particularly important as

*Work partially conducted while an intern at the Max Planck Institute for Security and Privacy, Germany

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

individuals are usually unaware of and cannot exercise agency over such relations and inferences since they are based on data shared by others Wachter and Mittelstadt (2019); Viljoen (2021).

However, existing algorithmic governance frameworks and prominent data importance and privacy approaches do not adequately account for algorithmically mediated user relations. Instead, they usually consider "data as an individual medium" Viljoen (2021) and treat users as independent from one another. Therefore, in this paper, taking as a case study CF algorithms where users are algorithmically interrelated by design, we ask the research question: *In practice, do we need to account for the algorithmically mediated relations created between users in a CF setting?* In answering this, we make the following contributions [2]:

Conceptual: we outline how legal and policy approaches and technical implementations account for algorithmically mediated user relations and discuss commonalities and discrepancies;
Empirical: we examine whether user independence can be assumed for two popular CF algorithms, namely user-based nearest neighbors (User kNN) and Simon Funk matrix factorization (Funk MF).

2 Conceptual Acknowledgement of Algorithmically Mediated User Relations

We provide a conceptual overview of how existing legal and policy frameworks and technical approaches account for algorithmically mediated relations and highlight tensions between disciplines.

Law & Technology Policy: *Data Privacy & Protection Laws* regulate how data about people are collected, processed, and used Parsons and Viljoen (2023); Viljoen (2021); Zuziak et al. (2023). Legal scholars have identified an individualistic focus of corresponding regimes, with collective interests and harms being mostly addressed by protecting the rights of affected individuals Mantelero (2017); Viljoen (2021). This individualistic focus has been considered inadequate within the context of predictive analytics because individuals cannot control how they relate to others Mantelero (2017); Viljoen (2021); Barocas and Levy (2020). Moreover, Wachter and Mittelstadt (2019) state that under the General Data Protection Regulation (GDPR), individuals are entitled to little control over the inferences that can be drawn about them. *Anti-discrimination Laws* do not necessarily protect user groups that result from algorithmic processing because their characteristics might not be straightforwardly associated with attributes of groups the law protects Wachter (2022).
Transparency Mandates & Individual Choice and Control Mechanisms in Personalization Services, such as the EU Digital Services Act (2022), the General Data Protection Regulation (2016), and the Filter Bubble Transparency Act (2021), have sought to address the challenges of algorithmic personalization by mandating that platforms should disclose details about data collection and algorithmic processing and allow users to exercise control over how their data are used. Such proposals, however, would likely fall short in addressing personalization-driven harms since the content that each user receives often depends on data shared by others Gordon-Tapiero et al. (2022); Viljoen (2021).

Technical Implementations: *Data Importance* approaches estimate the importance of training instances in a model's predictions and have been used for data and model debugging, data valuation, understanding model behavior, and detecting dataset errors. Methods can be categorized as leave-one-out and expected-improvement approaches Karlaš et al. (2022). *Leave-One-Out methods*, such as influence functions Koh and Liang (2017), estimate the importance of a training instance as the decrease in a quantity of interest when this point is removed from the training set. To estimate the effect of groups of training instances (e.g., groups of users), Koh et al. (2019) assume that instances are independent from one another. Under this assumption, influence functions have been used in CF recommenders to identify subsets of users primarily responsible for the recommendations that others receive Fang et al. (2020); Wu et al. (2021); Eskandanian et al. (2019). *Expected-Improvement Methods* capture interactions between subsets of training instances. They model the importance of a training instance by considering all the possible subsets of the training set and estimating the decrease in a quantity of interest when the training point is removed from all those possible subsets. Shapley values Shapley et al. (1953) have been used to quantify the importance of each training point Ghorbani and Zou (2019). However, to simplify the computation of Shapley values for groups of training instances, the groups are usually considered independent from one another Lundberg and Lee (2017).
Differential Privacy Dwork et al. (2014) has emerged as the criterion standard to provide privacy-preserving query answers over a statistical database by adding plausible deniability about the presence

[2] All data, code, and analysis are available on the project's Github repository.

of an individual record or group of records in the database. By assuming that the deletion of an individual record can eliminate all evidence of its participation in the data generation process, differential privacy mechanisms assume that data instances, and accordingly individuals, are independent from one another. However, deletion does not necessarily eliminate all of the record's evidence of participation in the data generation process, especially if the records are correlated Kifer and Machanavajjhala (2011, 2014); Gehrke et al. (2011). Moreover, applying group differential privacy in correlated tuples destroys all data utility Song et al. (2017). While Pufferfish privacy Kifer and Machanavajjhala (2014) has been developed to account for correlated data, there is currently no computationally efficient mechanism for it Song et al. (2017).

Commonalities and Discrepancies Based on prior works, algorithmically mediated user relations are not necessarily acknowledged in existing legal and policy approaches and prominent technical implementations for data importance and privacy. However, as discussed earlier, related literature has highlighted the need to account for these relations given that users might not be able to control them. This leaves open the question of whether accounting for algorithmically mediated relations is relevant in practice, which we empirically examine in the remainder of this paper.

3 Experimental Illustration in Recommender Systems

We empirically investigate whether in practice we need to account for the algorithmically mediated relations created between users in CF recommender systems. To do so, we examine whether users' influences on the predicted ratings of others' can be considered independent from one another. We adapt the definitions of Fang et al. (2020) and define the influences:
User-to-User: quantifies the influence of user's u data in the training set on user's v predicted ratings

$$I(u,v) \coloneqq \sum_{i \in I_u} \sum_{j \in I} |\hat{r}_{vj}(\theta^*_{D \setminus \{r_{ui}\}}) - \hat{r}_{vj}(\theta^*_D)| \tag{1}$$

Group-to-User (assuming independence of users in S): quantifies the influence of a group of users' S data in the training set on the predicted ratings of user v

$$I_{independence}(S,v) \coloneqq \sum_{u \in S} I(u,v) \tag{2}$$

Group-to-User (without assuming user independence): quantifies the influence of a group of users' S data in the training set on the predicted ratings of user v

$$I_{relations}(S,v) \coloneqq \sum_{j \in I} |\hat{r}_{vj}(\theta^*_{D \setminus R_S}) - \hat{r}_{vj}(\theta^*_D)| \tag{3}$$

where: r_{ui} user u's observed rating for item i, D the set of observed ratings (training set), I the available items, $I_u \subseteq I$ the items rated by u, and $R_S \subseteq D$ the interactions of users in set S. $D \setminus \{r_{ui}\}$ is D without r_{ui} and $D \setminus R_S$ is D without the interactions R_S. Finally, θ^*_D are the optimal model parameters given D and $\hat{r}_{vj(\cdot)}$ is user v's predicted rating for item j given parameters $\cdot$.

We examine whether:

$$I_{independence}(S,v) \simeq I_{relations}(S,v) \tag{4}$$

3.1 Setup

Our goal is to compute the difference in the predictions when subsets of users S are removed from the training set, when we consider 1) users in S independent from one another, and 2) algorithmically mediated relations between users in S. We aim to examine whether there are significant differences when user independence is assumed.

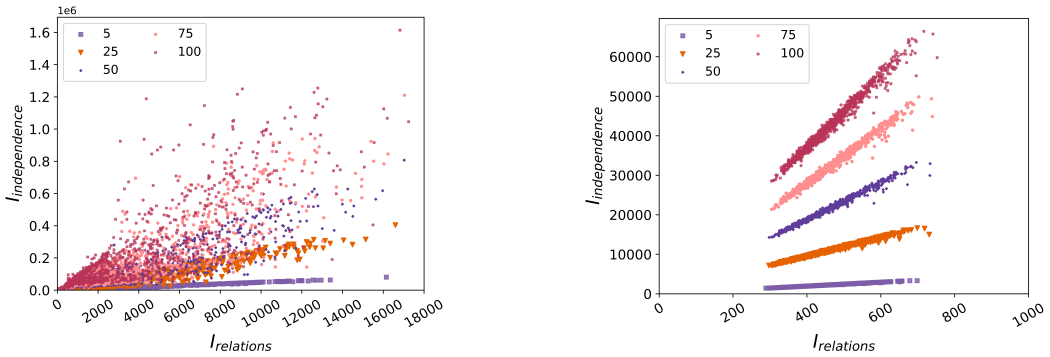
Group Construction Strategy: For each user v, we construct groups S of the most influential 5, 10, 25, 50, 75, and 100 users based on Equation 1 and the corresponding reduced training sets $D \setminus R_S$. We did not investigate the removal of larger groups since this could lead to changes in models' hyperparameters.

Influence Computation: For each S, v, we compute $I_{relations}$ 3 and $I_{independence}$ 2 for all available items. Despite the cost, we computed the actual differences in the predicted interactions by retraining the model with each reduced training set, instead of approximating them Koh et al. (2019).

Dataset & Algorithms: We run the experiment on MovieLens 100k where the number of users is relatively small. We examine Equation 4 for User kNN using the cosine similarity and Funk MF. For User kNN the selected k determines the size of a user's neighborhood for each predicted rating $\hat{r}_{ui}$. For Funk MF the number of latent factors determines the number of linearly independent users. Further details can be found in the supplementary material.

3.2 Results

Figure 1 illustrates the difference between $I_{independence}$ and $I_{relations}$ for User kNN and Funk MF on MovieLens 100k for the constructed user groups S of varying sizes.



(a) User kNN, MovieLens 100k

(b) Funk MF, MovieLens 100k

Figure 1: Analysis of user independence for the User kNN (left), and Funk MF (right) algorithms on the MovieLens 100k dataset. User independence might be inadequate for User kNN (left), while for Funk MF (right) users can be approximately considered as independent from one another.

For User kNN and for all users and constructed groups, we observe that there is a significant difference between the actual influence of removing a group and the influence value of a group under the user independence assumption (Sapiro-Wilk-Test: p « 0.05, Wilcoxon-Test: p « 0.05). The correlation between the values of $I_{independence}$ and $I_{relations}$ substantially decreases as the size of S increases (Table 5 in supplementary material).

For Funk MF, there is a significant difference between the values of $I_{relations}$ and $I_{independence}$ (Sapiro-Wilk-Test: p « 0.05, Wilcoxon-Test: p « 0.05). However, despite the significant magnitude difference, there is high correlation between the values (Table 5 in supplementary material).

Accordingly, assuming user independence in practice depends on the CF algorithm and the size of the set of users whose influence we want to compute. Given that individual model hyperparameters, such as the number of nearest neighbors for User kNN (or the latent factors for Funk MF), affect the considered degree of similarity (or linear correlation) between users' predictions, further investigation on their impact on algorithmically mediated user relations is planned in future work.

4 Conclusion & Broader Impact

In this paper, we investigate the practical relevancy of algorithmically mediated user relations. We discuss the importance of acknowledging such relations in personalization services and show that existing algorithmic governance frameworks and technical approaches do not necessarily account for them. Instead, they consider users as independent from one another. Given this conceptual discrepancy, we empirically examine whether accounting for algorithmically mediated user relations is relevant in practice within the context of CF recommenders, which by design relate users. Our results show that depending on the algorithm, assuming user independence might not be adequate.

As future work, we plan to study the effect of model hyperparameters on algorithmically mediated user relations. We believe that further understanding of these relations will provide insights to practitioners on related problems such as data deletion Ginart et al. (2019), data minimization Shanmugam et al. (2022), and data influence Koh et al. (2019), as well as to technology policy stakeholders to revise and design existing and future algorithmic and data governance frameworks.

References

Charu C Aggarwal et al. 2016. *Recommender Systems: The Textbook*. Vol. 1. Springer.

Solon Barocas and Karen Levy. 2020. Privacy dependencies. *Washington Law Review* 95 (2020), 555.

Emmanuel Candes and Benjamin Recht. 2012. Exact matrix completion via convex optimization. *Commun. ACM* 55, 6 (2012), 111–119.

Yuejie Chi. 2018. Low-rank matrix completion [lecture notes]. *IEEE Signal Processing Magazine* 35, 5 (2018), 178–181.

Jin Yao Chin, Yile Chen, and Gao Cong. 2022. The Datasets Dilemma: How Much Do We Really Know About Recommendation Datasets?. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 141–149.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

Farzad Eskandanian, Nasim Sonboli, and Bamshad Mobasher. 2019. Power of the Few: Analyzing the Impact of Influential Users in Collaborative Recommender Systems. 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, 225–233. https://doi.org/10.1145/3320435.3320464

EU Digital Services Act. 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065

Minghong Fang, Neil Zhenqiang Gong, and Jia Liu. 2020. Influence function based data poisoning attacks to top-n recommender systems. In *Proceedings of The Web Conference 2020*. 3019–3025.

Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–49.

Filter Bubble Transparency Act. 2021. https://www.congress.gov/bill/117th-congress/senate-bill/2024/text?format=txt

Simon Funk. 2006. https://sifter.org/simon/journal/20061211.html

Johannes Gehrke, Edward Lui, and Rafael Pass. 2011. Towards privacy for social networks: A zero-knowledge based definition of privacy. In *Theory of Cryptography Conference*. Springer, 432–449.

General Data Protection Regulation. 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*. PMLR, 2242–2251.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making AI Forget You: Data Deletion in Machine Learning. *Advances in Neural Information Processing Systems* 32 (2019).

Ayelet Gordon-Tapiero, Alexandra Wood, and Katrina Ligett. 2022. The Case for Establishing a Collective Perspective to Address the Harms of Platform Personalization. In *Proceedings of the 2022 Symposium on Computer Science and Law* (Washington DC, USA) *(CSLAW '22)*. Association for Computing Machinery, New York, NY, USA, 119–130. https://doi.org/10.1145/3511265.3550450

Mathias Humbert, Benjamin Trubert, and Kévin Huguenin. 2019. A survey on interdependent privacy. *ACM Computing Surveys (CSUR)* 52, 6 (2019), 1–40.

Lanah Kammourieh, Thomas Baar, Jos Berens, Emmanuel Letouzé, Julia Manske, John Palmer, David Sangokoya, and Patrick Vinck. 2017. *Group Privacy in the Age of Big Data*. Springer Nature, Gewerbestrasse 11, 6330 Cham, Switzerland, Chapter 3, 37–66. https://doi.org/10.1007/978-3-319-46608-8_3

Bojan Karlaš, David Dao, Matteo Interlandi, Bo Li, Sebastian Schelter, Wentao Wu, and Ce Zhang. 2022. Data debugging with shapley importance over end-to-end machine learning pipelines. *arXiv preprint arXiv:2204.11131* (2022).

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the Spread of Influence through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, D.C.) *(KDD '03)*. Association for Computing Machinery, New York, NY, USA, 137–146. https://doi.org/10.1145/956750.956769

Daniel Kifer and Ashwin Machanavajjhala. 2011. No Free Lunch in Data Privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data* (Athens, Greece) *(SIGMOD '11)*. Association for Computing Machinery, New York, NY, USA, 193–204. https://doi.org/10.1145/1989323.1989345

Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A Framework for Mathematical Privacy Definitions. *ACM Transactions of Database Systems* 39, 1, Article 3 (jan 2014), 36 pages. https://doi.org/10.1145/2514689

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*. PMLR, 1885–1894.

Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. 2019. On the accuracy of influence functions for measuring group effects. *Advances in Neural Information Processing Systems* 32.

Yehuda Koren, Steffen Rendle, and Robert Bell. 2021. Advances in collaborative filtering. *Recommender Systems Handbook* (2021), 91–142.

Qian Li, Xiangmeng Wang, Zhichao Wang, and Guandong Xu. 2023. Be Causal: De-Biasing Social Network Confounding in Recommendation. *ACM Transactions on Knowledge Discovery from Data* 17, 1, Article 14 (feb 2023), 23 pages. https://doi.org/10.1145/3533725

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

Alessandro Mantelero. 2017. From group privacy to collective privacy: towards a new dimension of privacy and data protection in the big data era. *Group Privacy: New Challenges of Data Technologies* (2017), 139–158.

Amanda Parsons and Salomé Viljoen. 2023. Valuing Social Data. *University of Colorado Law Legal Studies Research Paper* 23-16 (2023).

Divya Shanmugam, Fernando Diaz, Samira Shabanian, Michèle Finck, and Asia Biega. 2022. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 839–849.

Lloyd S Shapley et al. 1953. A value for n-person games. (1953).

Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. 2017. Pufferfish Privacy Mechanisms for Correlated Data. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) *(SIGMOD '17)*. Association for Computing Machinery, New York, NY, USA, 1291–1306. https://doi.org/10.1145/3035918.3064025

Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Geng. 2020. Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 23–32.

Salome Viljoen. 2021. A relational theory of data governance. *The Yale Law Journal* 131 (2021), 573.

Sandra Wachter. 2022. The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law. *Tulane Law Review* 97 (2022), 149.

Sandra Wachter and Brent Mittelstadt. 2019. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review* (2019), 494.

Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, and Enhong Chen. 2021. Triple adversarial learning for influence based poisoning attack in recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1830–1840.

Maciej Krzysztof Zuziak, Onntje Hinrichs, Aizhan Abdrassulova, and Salvatore Rinzivillo. 2023. Data Collaboratives with the Use of Decentralised Learning. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 615–625.

5 Supplementary Material

5.1 Datasets

We report the values of dataset metrics suggested by Chin et al. (2022).

Table 1: Dataset Metrics

Dataset	$\#Users$	$\#Items$	$\#Interactions$	$Space_{log_{10}}$	$Shape_{log_{10}}$	$Density_{log_{10}}$	$Gini_u$	$Gini_i$
MovieLens 100k	6,040	3,952	1,000,209	4.378	0.184	-1.378	0.33	0.339

We loaded the MovieLens 100k dataset from the Microsoft Recommenders [3] package. The dataset is pre-processed and all users have at least 20 interactions. Since there is no clearly advantageous method to split a dataset into training and test sets, we follow the majority practice of global split-by-ratio Sun et al. (2020) and split each dataset into a training and test by ratio $80:20$. The split is performed randomly but according to a stratified split to maintain local per-user ratios and ensure the presence of all users in both sets. The training set was further split randomly to training and validation set in ratio $80:20$. We set the random seed to 42.

5.2 Algorithms

5.2.1 Overview of the Considered Recommendation Algorithms

Neighborhood-based Algorithms Neighborhood-based methods provide recommendations to a target user by identifying (i) users with similar item rating behavior (user-based) or (ii) similar items to the ones that the target user has interacted with based on the ratings of other users (item-based). Similarity can be based on a pre-defined metric or learned from the available data. For example, for the user-based k-nearest-neighbors algorithm (User kNN), a missing rating $\hat{r_{ui}}$ of user u for item i Koren et al. (2021); Aggarwal et al. (2016) is approximated by:

$$r_{ui} \simeq \hat{r}_{ui} = \frac{\sum_{v \in N_i^k(u)} sim(u,v) \cdot r_{vi}}{\sum_{v \in N_i^k(u)} sim(u,v)} \tag{5}$$

where $N_i^k(u)$ is the set of k users most similar to user u (by the value of a similarity measure $sim(u,v)$) who have rated item i. Therefore, by design, the users $N_i^k(u)$ are the ones related to u for the prediction $\hat{r}_{ui}$.

Low-rank Matrix Completion Algorithms The underlying assumption of such algorithms is that users' interactions are only affected by a few factors Koren et al. (2021); Aggarwal et al. (2016). In mathematical terms, it is assumed that there is a high linear correlation between the user-item interactions and, therefore, the interaction matrix R is low rank $k \ll min\{n,m\}$ Candes and Recht (2012); Chi (2018). Hence, the assumed data redundancies make it possible to construct a fully specified low-rank approximation of R, even when the number of observed interactions is small.

One family of low-rank matrix completion algorithms are matrix factorization (MF) techniques. Such techniques factorize R and estimate the unobserved ratings via the dot-product of two embedding vectors of dimensionality k. The simplest matrix factorization algorithm commonly used as a baseline is the one suggested by Simon Funk (Funk MF) Funk (2006). Written in matrix form, the interaction matrix for Funk MF is approximated by:

$$R \simeq \hat{R} = PQ^T \tag{6}$$

where $P \in \mathbb{R}^{n \times k}$ and $Q \in \mathbb{R}^{m \times k}$ are embedding matrices. The matrices P and Q are determined by minimizing a non-convex optimization function, set according to a desired objective (e.g., $min_{q,p} \sum_{u,i} (r_{ui} - q_i^T p_i)^2 - \lambda(||q_i||^2 + ||p_u||^2)$ Koren et al. (2021); Chi (2018). Irrespective of the objective to be optimized, the rank of $\tilde{R}$ is:

$$rank(\hat{R})_{Funk\_MF} = rank(PQ^T) \leq min(rank(P), rank(Q)) \leq k \tag{7}$$

Therefore, the number of linearly independent users based on their predicted ratings is at most k. Hence, this approach assumes that there are at least $(n-k)$ correlated users. Consequently, their predicted ratings are also correlated.

[3] Microsoft Recommenders

5.2.2 Implementation & Tuning Details

For reprodicibility purposes, we used the implementations of Ferrari Dacrema et al. (2021). We assume that items that a user has interacted with could be recommended again. Since we do not aim to draw conclusions about a specific dataset-algorithm setup, this is not a limiting factor.

For hyperparameter tuning, we used Bayesian search and tuned on NDCG@10. We set the random seed to 42. When evaluating, we used for each user 99 randomly picked negative (non-interacted) samples per positive sample for ranking along with the validation set. As suggested by Ferrari Dacrema et al. (2021), for all models we selected the number of epochs via early stopping. The hyperparameter search space and the obtained optimal hyperparameters are reported on Tables 2 and 3 respectively.

Table 2: Hyperparameter Search Space

Algorithm	Hyperparameter Space
User kNN (cosine)	topK: Integer(5,1000), shrink: Integer(0,1000), normalize: Categorical([True, False]), feature_weighting: Categorical([none, TF-IDF, BM25])
Funk MF	sgd_mode: Categorical([sgd, adagrad, adam]), num_factors: Integer(1, 200), epochs: Categorical([500]), use_bias: Categorical([False]), batch_size: Categorical([1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024]), item_reg: Real(low = 1e-5, high = 1e-2, prior = 'log-uniform'), user_reg: Real(low = 1e-5, high = 1e-2, prior = 'log-uniform'), learning_rate: Real(low = 1e-4, high = 1e-1, prior = 'log-uniform'), negative_interactions_quota: Real(low = 0.0, high = 0.5, prior = 'uniform')

Table 3: Optimal Hyperparameters per Algorithm and Dataset

Algorithm	Dataset	Optimal Hyperparameters
User kNN (cosine)	MovieLens 100k	topK: 185, shrink: 0, normalize: True, feature_weighting: none
Funk MF	MovieLens 100k	sgd_mode: adagrad, epochs: 365, use_bias: False, batch_size: 1, num_factors: 195, item_reg: 4.992453416923983e-05, user_reg: 1.8699039697141504e-05, learning_rate: 0.007164040428191017, negative_interactions_quota: 0.19123099563358142

5.3 Experimental Details

5.3.1 Group Construction Strategy

For each user u in the training set, we construct groups of the most influential 5, 10, 25, 50, 75, and 100 users based on the user-to-user influence relation, as given in Equation 1. The number of ratings corresponding to each group range is up to 14%. We did not investigate the removal of larger groups because we expected this to lead to changes in the model's hyperparameters. Table 4 summarizes the total number of ratings per group size.

Table 4: Percentage of Ratings Per Group Size of Influential Users in the Training Set of MovieLens 100k Dataset

$\#Users$	$\%Ratings$ User kNN	$\%Ratings$ Funk MF
5	$0.5 - 1.5\%$	$\leq 0.5 - 1.5\%$
25	$1.0 - 4.5\%$	$\leq 1.0 - 4.5\%$
50	$3.5 - 8.0\%$	$\leq 3.5 - 8.0\%$
75	$6.0 - 10.0\%$	$\leq 5.5 - 10.5\%$
100	$7.5 - 14\%$	$\leq 7.5 - 13.5\%$

5.3.2 Correlation between $I_{relations}$ and $I_{independence}$

The Kendall rank correlation coefficients are summarized in Table 5.

Table 5: Kendall's τ-b Coefficients between $I_{relations}$ and $I_{independence}$

Group Size	User kNN (cosine)	Funk MF
5	$\tau = 0.904$, $p < 0.05$	$\tau = 0.978$, $p < 0.05$
25	$\tau = 0.751$, $p \ll 0.05$	$\tau = 0.946$, $p < 0.05$
50	$\tau = 0.700$, $p \ll 0.05$	$\tau = 0.935$, $p < 0.05$
75	$\tau = 0.673$, $p \ll 0.05$	$\tau = 0.927$, $p < 0.05$
100	$\tau = 0.656$, $p \ll 0.05$	$\tau = 0.918$, $p < 0.05$
Overall	$\tau = 0.526$, $p < 0.05$	$\tau = 0.285$, $p \ll 0.05$