Robustness of Meta Matrix Factorization Against Strict Privacy Constraints

Peter Muellner[1(✉)], Dominik Kowald[1], and Elisabeth Lex[2]

[1] Know-Center GmbH, Graz, Austria
{pmuellner,dkowald}@know-center.at
[2] Graz University of Technology, Graz, Austria
elisabeth.lex@tugraz.at

Abstract. In this paper, we explore the reproducibility of MetaMF, a meta matrix factorization framework introduced by Lin et al. MetaMF employs meta learning for federated rating prediction to preserve users' privacy. We reproduce the experiments of Lin et al. on five datasets, i.e., Douban, Hetrec-MovieLens, MovieLens 1M, Ciao, and Jester. Also, we study the impact of meta learning on the accuracy of MetaMF's recommendations. Furthermore, in our work, we acknowledge that users may have different tolerances for revealing information about themselves. Hence, in a second strand of experiments, we investigate the robustness of MetaMF against strict privacy constraints. Our study illustrates that we can reproduce most of Lin et al.'s results. Plus, we provide strong evidence that meta learning is essential for MetaMF's robustness against strict privacy constraints.

Keywords: Recommender systems · Privacy · Meta learning · Federated learning · Reproducibility · Matrix factorization

1 Introduction

State-of-the-art recommender systems learn a user model from user and item data and the user's interactions with items to generate personalized recommendations. In that process, however, users' personal information may be exposed, resulting in severe privacy threats. As a remedy, recent research makes use of techniques like federated learning [2,4,6] or meta learning [7,20] to ensure privacy in recommender systems. In the federated learning paradigm, no data ever leaves a user's device, and as such, the leakage of their data by other parties is prohibited. With meta learning, a model gains the ability to form its hypothesis based on a minimal amount of data.

Similar to recent work [5,15], MetaMF by Lin et al. [16] combines federated learning with meta learning to provide personalization and privacy. Besides, MetaMF exploits collaborative information among users and distributes a private rating prediction model to each user. Due to MetaMF's recency and its clear focus on increasing privacy for users via a novel framework, we are interested

in the reproducibility of Lin et al.'s research. Additionally, we aim to contribute our own branch of research regarding privacy, i.e., MetaMF's robustness against strict privacy constraints. This is motivated by a statement of Lin et al. about one critical limitation of MetaMF, i.e., its sensitivity to data scarcity that could arise when users employ strict privacy constraints by withholding a certain amount of their data. In this regard, every user has a certain privacy budget, i.e., a budget of private data she is willing to share. Thus, in our paper at hand, the privacy budget is considered a measure of how much data disclosure a user tolerates and is defined as the fraction of rating data she is willing to share with others. Thereby, employing small privacy budgets and thus, withholding data, serves as a realization of strict privacy constraints.

Our work addresses MetaMF's limitation against data scarcity and is structured in two parts. First, we conduct a study with the aim to reproduce the results given in the original work by Lin et al. Concretely, we investigate two leading research questions, i.e., *RQ1a: How does MetaMF perform on a broad body of datasets?* and *RQ1b: What evidence does MetaMF provide for personalization and collaboration?* Second, we present a privacy-focused study, in which we evaluate the impact of MetaMF's meta learning component and test MetaMF's performance on users with different amounts of rating data. Here, we investigate two more research questions, i.e., *RQ2a: What is the role of meta learning in the robustness of MetaMF against decreasing privacy budgets?* and *RQ2b: How do limited privacy budgets affect users with different amounts of rating data?* We address *RQ1a* and *RQ1b* in Sect. 3 by testing MetaMF's predictive capabilities on five different datasets, i.e., Douban, Hetrec-MovieLens, MovieLens 1M, Ciao, and Jester. Here, we find that most results provided by Lin et al. can be reproduced. In Sect. 4, we elaborate on *RQ2a* and *RQ2b* by examining MetaMF in the setting of decreasing privacy budgets. Here, we provide strong evidence of the important role of meta learning in MetaMF's robustness. Besides, we find that users with large amounts of rating data are substantially disadvantaged by decreasing privacy budgets compared to users with few rating data.

2   Methodology

In this section, we illustrate our methodology of addressing *RQ1a* and *RQ1b*, i.e., the reproducibility of Lin et al. [16], and *RQ2a* and *RQ2b*, i.e., MetaMF's robustness against decreasing privacy budgets.

2.1   Approach

MetaMF. Lin et al. recently introduced a novel matrix factorization framework in a federated environment leveraging meta learning. Their framework comprises three steps. First, collaborative information among users is collected and subsequently, utilized to construct a user's collaborative vector. This collaborative vector serves as basis of the second step. Here, in detail, the parameters of

a private rating prediction model are learned via meta learning. Plus, in parallel, personalized item embeddings, representing a user's personal "opinion" about the items, are computed. Finally, in the third step, the rating of an item is predicted utilizing the previously learned rating prediction model and item embeddings. We resort to MetaMF to address *RQ1a*, *RQ1b*, and *RQ2b*, i.e., the reproducibility of results presented by Lin et al. and the influence of decreasing privacy budgets on users with different amounts of rating data.

NoMetaMF. In our privacy-focused study, *RQ2a* addresses the role of meta learning in MetaMF's robustness against decreasing privacy budgets. Thus, we conduct experiments with and without MetaMF's meta learning component. For the latter kind of experiments, we introduce NoMetaMF, a variant of MetaMF with no meta learning. In MetaMF, a private rating prediction model is generated for each user by leveraging meta learning. The authors utilize a hypernetwork [11], i.e., a neural network, coined meta network, that generates the parameters of another neural network. Based on the user's collaborative vector $\mathbf{c}_u$, the meta network generates the parameters of the rating prediction model, i.e., weights $\mathbf{W}_l^u$ and biases $\mathbf{b}_l^u$ for layer l and user u. This is given by

$$\mathbf{h} = \mathrm{ReLU}(\mathbf{W}_h^* \mathbf{c}_u + \mathbf{b}_h^*) \tag{1}$$

$$\mathbf{W}_l^u = \mathbf{U}_{W_l^u}^* \mathbf{h} + \mathbf{b}_{W_l^u}^* \tag{2}$$

$$\mathbf{b}_l^u = \mathbf{U}_{b_l^u}^* \mathbf{h} + \mathbf{b}_{b_l^u}^* \tag{3}$$

where $\mathbf{h}$ is the hidden state with the widely-used $\mathrm{ReLU}(x) = \max(0, x)$ [8,12] activation function, $\mathbf{W}_h^*$, $\mathbf{U}_{W_l^u}^*$, $\mathbf{U}_{b_l^u}^*$ are the weights and $\mathbf{b}_h^*$, $\mathbf{b}_{W_l^u}^*$, $\mathbf{b}_{b_l^u}^*$ are the biases of the meta network. NoMetaMF excludes meta learning by disabling backpropagation through the meta network in Eqs. 1–3. Thus, meta parameters $\mathbf{W}_h^*$, $\mathbf{U}_{W_l^u}^*$, $\mathbf{U}_{b_l^u}^*$, $\mathbf{b}_h^*$, $\mathbf{b}_{W_l^u}^*$, $\mathbf{b}_{b_l^u}^*$ will not be learned in NoMetaMF. While backpropagation is disabled in the meta network, parameters W_l^u and b_l^u are learned over those non-meta parameters in NoMetaMF to obtain the collaborative vector. Hence, the parameters of the rating prediction models are still learned for each user individually, but without meta learning.

Lin et al. also introduce a variant of MetaMF, called MetaMF-SM, which should not be confused with NoMetaMF. In contrast to MetaMF, MetaMF-SM does not generate a private rating prediction model for each user individually, but instead utilizes a shared rating prediction model for all users. Our NoMetaMF model generates an individual rating prediction model for each user but operates without meta learning. Furthermore, we note that in our implementation of NoMetaMF, the item embeddings are generated in the same way as in MetaMF. With NoMetaMF, we aim to investigate the impact of meta learning on the robustness of MetaMF against decreasing privacy budgets, i.e., *RQ2a*.

2.2 Datasets

In line with Lin et al., we conduct experiments on four datasets: Douban [14], Hetrec-MovieLens [3], MovieLens 1M [13], and Ciao [10]. We observe that none

of these datasets comprises a high average number of ratings per item, i.e., 22.6 (Douban), 85.6 (Hetrec-MovieLens), 269.8 (MovieLens 1M), and 2.7 (Ciao). To increase the diversity of our datasets, we include a fifth dataset to our study, i.e., Jester [9] with an average number of ratings per item of 41,363.6. Furthermore, Lin et al. claimed that several observations about Ciao may be explained by its low average number of ratings per user, i.e., 38.3. Since Jester exhibits a similarly low average number of ratings per user, i.e., 56.3, we utilize Jester to verify Lin et al.'s claims. To fit the rating scale of the other datasets, we scale Jester's ratings to a range of [1, 5]. Descriptive statistics of our five datasets are outlined in detail in the following lines. *Douban* comprises 2,509 users with 893,575 ratings for 39,576 items. *Hetrec-MovieLens* includes 10,109 items and 855,598 ratings of 2,113 users. The popular *MovieLens 1M* dataset includes 6,040 users, 3,706 items and 1,000,209 ratings. *Ciao* represents 105,096 items, with 282,619 ratings from 7,373 users. Finally, our additional *Jester* dataset comprises 4,136,360 ratings for 100 items from 73,421 users.

We follow the evaluation protocol of Lin et al. and thus, perform no cross-validation. Therefore, each dataset is randomly separated into 80% training set R_{train}, 10% validation set R_{val} and 10% test set R_{test}. However, we highlight that in the case of Douban, Hetrec-MovieLens, MovieLens 1M, and Ciao, we utilize the training, validation and test set provided by Lin et al.

Identification of User Groups. In *RQ2b*, we study how decreasing privacy budgets influence the recommendation accuracy of user groups with different user behavior. That is motivated by recent research [1,19], which illustrates differences in recommendation quality for user groups with different characteristics. As an example, [19] measures a user group's mainstreaminess, i.e., how the user groups' most listened artists match the most listened artists of the entire population. The authors split the population into three groups of users with low, medium, and high mainstreaminess, respectively. Their results suggest that low mainstream users receive far worse recommendations than mainstream users.

In a similar vein, we also split users into three user groups: *Low*, *Med*, and *High*, referring to users with a low, medium, and a high number of ratings, respectively. To precisely study the effects of decreasing privacy budgets on each user group, we generate them such that the variance of the number of ratings is low, but yet, include a sufficiently large number of users. For this matter, each of our three user groups includes 5% of all users. In detail, we utilize the 5% of users with the least ratings (i.e., *Low*), the 5% of users with the most ratings (i.e., *High*) and the 5% of users, whose number of ratings are the closest to the median (i.e., *Med*). Thus, each user group consists of 125 (Douban), 106 (Hetrec-MovieLens), 302 (MovieLens 1M), 369 (Ciao), and 3,671 (Jester) users.

2.3   Recommendation Evaluation

In concordance to the methodology of Lin et al., we minimize the mean squared error (MSE) between the predicted $\hat{r} \in \hat{R}$ and the real ratings $r \in R$ as the

objective function for training the model. Additionally, we report the MSE and the mean absolute error (MAE) on the test set R_{test} to estimate our models' predictive capabilities. Since we dedicate parts of this work to shed light on MetaMF's and NoMetaMF's performance in settings with different degrees of privacy, we illustrate how we simulate decreasing privacy budgets and how we evaluate a model's robustness against these privacy constraints.

Simulating Different Privacy Budgets. To simulate the reluctance of users to share their data, we propose a simple sampling procedure in Algorithm 1. Let β be the privacy budget, i.e., the fraction of data to be shared. First, a user u randomly selects a fraction of β of her ratings without replacement. Second, the random selection of ratings R_u^β is then shared by adding it to the set R^β. That ensures that (i) each user has the same privacy budget β and (ii) each user shares at least one rating to receive recommendations. The set of shared ratings R^β without held back ratings then serves as a training set for our models.

Algorithm 1: Sampling procedure for simulating privacy budget β.

Input: Ratings R, Users U and privacy budget β.
Result: Shared ratings R^β, with a fraction of β of each user's ratings.
$R^\beta = \{\}$
for $u \in U$ **do**
  $R_u^\beta = \{R_u' \subseteq R_u : |R_u'|/|R_u| = \beta\}$
  $R^\beta = R^\beta \cup R_u^\beta$
end

Measuring Robustness. Our privacy-focused study is concerned with discussing MetaMF's robustness against decreasing privacy budgets. We quantify a model's robustness by how the model's predictive capabilities change by decreasing privacy budgets. In detail, we introduce a novel accuracy measurement called ΔMAE@β, which is a simple variant of the mean absolute error.

Definition 1 (ΔMAE@β). *The relative mean absolute error ΔMAE@β measures the predictive capabilities of a model M under a privacy budget β relative to the predictive capabilities of M without any privacy constraints.*

$$\text{MAE@}\beta = \frac{1}{|R_{test}|} \sum_{r_{u,i} \in R_{test}} |(r_{u,i} - M(R_{train}^\beta, \theta)_{u,i})| \qquad (4)$$

$$\Delta\text{MAE@}\beta = \frac{\text{MAE@}\beta}{\text{MAE@}1.0} \qquad (5)$$

where $M(R_{train}^\beta, \theta)_{u,i}$ is the estimated rating for user u on item i for M with parameters θ being trained on the dataset R_{train}^β and $|\cdot|$ is the absolute function. Please note that the same R_{test} is utilized for different values of β.

Table 1. MetaMF's error measurements (reproduced/original) for our five datasets alongside the MAE (mean absolute error) and the MSE (mean squared error) reported in the original paper. The non-reproducibility of the MSE on the Ciao dataset can be explained by the particularities of the MSE and the Ciao dataset. All other measurements can be reproduced (*RQ1a*).

Dataset	MAE	MSE
Douban	0.588/0.584	0.554/0.549
Hetrec-MovieLens	0.577/0.571	0.587/0.578
MovieLens 1M	0.687/0.687	0.765/0.760
Ciao	0.774/0.774	1.125/1.043
Jester	0.856/-	1.105/-

Furthermore, it is noteworthy that the magnitude of ΔMAE@β measurements does not depend on the underlying dataset, as it is a relative measure. Thus, one can compare a model's ΔMAE@β measurements among different datasets.

2.4   Source Code and Materials

For the reproducibility study, we utilize and extend the original implementation of MetaMF, which is provided by the authors alongside the Douban, Hetrec-MovieLens, MovieLens 1M, and Ciao dataset samples via BitBucket[1]. Furthermore, we publish the entire Python-based implementation of our work on GitHub[2] and our three user groups for all five datasets on Zenodo[3] [18].

We want to highlight that we are not interested in outperforming any state-of-the-art approaches on our five datasets. Thus, we refrain from conducting any hyperparameter tuning or parameter search and utilize precisely the same parameters, hyperparameters, and optimization algorithms as Lin et al. [16].

3   Reproducibility Study

In this section, we address *RQ1a* and *RQ1b*. As such, we repeat experiments by Lin et al. [16] to verify the reproducibility of their results. Therefore, we evaluate MetaMF on the four datasets Douban, Hetrec-MovieLens, MovieLens 1M, and Ciao. Additionally, we measure its accuracy on the Jester dataset. Please note that we strictly follow the evaluation procedure as in the work to be reproduced.

We provide MAE (mean absolute error) and MSE (mean squared error) measurements on our five datasets in Table 1. It can be observed that we can reproduce the results by Lin et al. up to a margin of error smaller than 2%. Only in

---

[1] https://bitbucket.org/HeavenDog/metamf/src/master/, Last accessed Oct. 2020.
[2] https://github.com/pmuellner/RobustnessOfMetaMF.
[3] https://doi.org/10.5281/zenodo.4031011.

the case of the MSE on the Ciao dataset, we obtain different results. Due to the selection of random batches during training, our model slightly deviates from the one utilized by Lin et al. Thereby, also, the predictions are likely to differ marginally. As described in [21], the MSE is much more sensitive to the variance of the observations than the MAE. Thus, we argue that the non-reproducibility of the MSE on the Ciao dataset can be explained by the sensitivity of the MSE on the variance of the observations in each batch. In detail, we observed in Sect. 2.2 that Ciao comprises very few ratings but lots of items. Thus, the predicted ratings are sensitive to the random selection of training data within each batch. However, it is noteworthy that we can reproduce the more stable MAE on the Ciao dataset. Hence, we conclude that our results provide strong evidence of the originally reported measurements being reproducible, enabling us to answer *RQ1a* in the affirmative.

Next, we study the rating prediction models' weights and the learned item embeddings. Again, we follow the procedure of Lin et al. and utilize the popular t-SNE (t-distributed stochastic neighborhood embedding) [17] method to reduce the dimensionality of the weights and the item embeddings to two dimensions. Since Lin et al. did not report any parameter values for t-SNE, we rely on the default parameters, i.e., we set the perplexity to 30 [17]. After the dimensionality reduction, we standardize all observations $x \in X$ by $\frac{x-\mu}{\sigma}$, where μ is the mean and σ is the standard deviation of X. The rating prediction model of each user is defined as a two-layer neural network. However, we observe that Lin et al. did not describe what layer's weights they visualize. Correspondences with the leading author of Lin et al. clarified that in their work, they only describe the weights of the first layer of the rating prediction models. The visualizations of the first layer's weights of the rating prediction models on our five datasets are given in Fig. 1.

In line with Lin et al., we discuss the weights and the item embeddings with respect to personalization and collaboration. As the authors suggest, personalization leads to distinct weight embeddings and collaboration leads to clusters within the embedding space. First, we observe that MetaMF tends to generate different weight embeddings for each user. Second, the visualizations exhibits well-defined clusters, which indicates that MetaMF can exploit collaborative information among users. However, our visualizations of the weights deviate slightly from the ones reported by Lin et al. Similar to the reproduction of the accuracy measurements in Table 1, we attribute this to the inability to derive the exact same model as Lin et al. Besides, t-SNE comprises random components and thus, generates slightly varying visualizations. However, the weights for the Ciao dataset in Fig. 1d illustrate behavior that contradicts Lin et al.'s observations. In the case of the Ciao dataset, they did not observe any form of clustering and attributed this behavior to the small number of ratings per user in the Ciao dataset. To test their claim, we also illustrate the Jester dataset with a similarly low number of ratings per user. In contrast, our visualizations indeed show well-defined clusters and different embeddings. We note that Jester exhibits many more clusters than the other datasets due to the much larger

(a) Douban   (b) Hetrec-MovieLens   (c) MovieLens 1M

(d) Ciao   (e) Jester
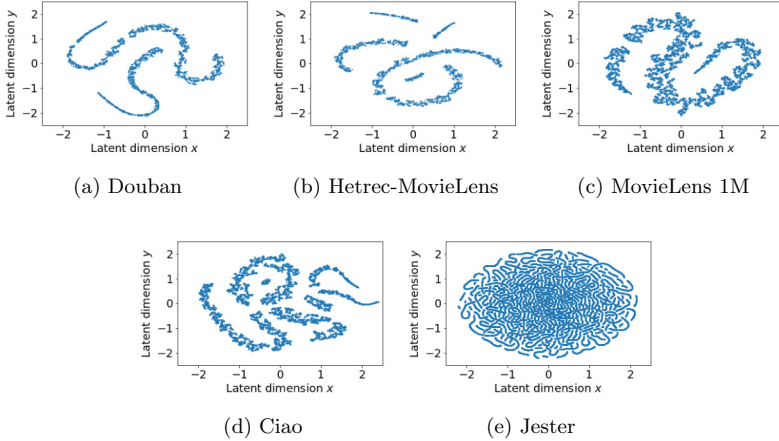
Fig. 1. MetaMF's weights embeddings of the first layer of the rating prediction models. One observation corresponds to an individual user ($RQ1b$).

number of users. Overall, we find that both, Ciao and Jester, do not support the claim made by Lin et al. However, we see the possibility that this observation may be caused by randomness during training.

Due to space limitations, we refrain from visualizing the item embeddings. It is worth noticing that our observations on the weights also hold for the item embeddings. In detail, our visualizations exhibit indications of collaboration and personalization for all datasets. Overall, we find the visualizations of the weights and the item embeddings presented by Lin et al. to be reproducible for the Douban, Hetrec-MovieLens, and MovieLens 1M datasets and thus, we can also positively answer $RQ1b$.

4   Privacy-Focused Study

In the following, we present experiments that go beyond reproducing Lin et al.'s work [16]. Concretely, we explore the robustness of MetaMF against decreasing privacy budgets and discuss $RQ2a$ and $RQ2b$. More detailed, we shed light on the effect of decreasing privacy budgets on MetaMF in two settings: (i) the role of MetaMF's meta learning component and (ii) MetaMF's ability to serve users with different amounts of rating data equally well.

First, we compare MetaMF to NoMetaMF in the setting of decreasing privacy budgets. Therefore, we utilize our sampling procedure in Algorithm 1 to generate datasets with different privacy budgets. In detail, we construct 10 training sets, i.e., $\{R_{train}^{\beta} : \beta \in \{1.0, 0.9, \ldots, 0.2, 0.1\}\}$, on which MetaMF and NoMetaMF are trained on. Then, we evaluate both models on the test set R_{test}. It is worth noticing that R_{test} is the same for all values of β to enable a valid comparison. Our results in Fig. 2a illustrate that for all datasets, MetaMF preserves its
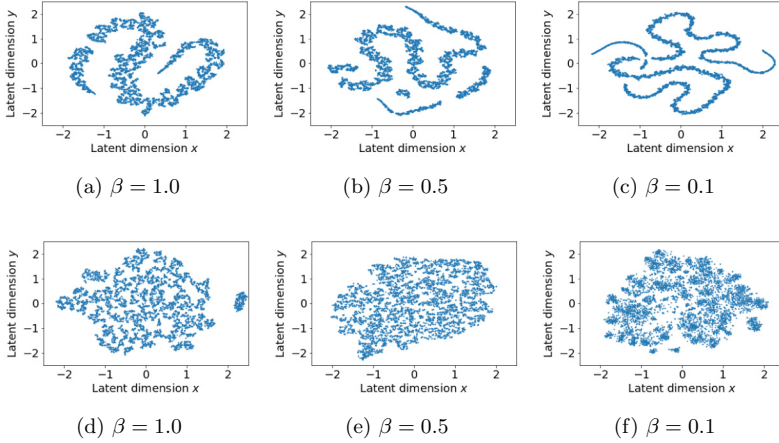
(a) MetaMF   (b) NoMetaMF
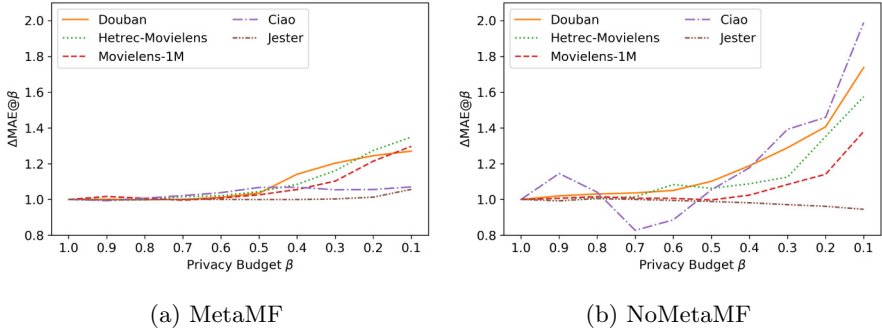
Fig. 2. ΔMAE@β measurements on (a) MetaMF and (b) NoMetaMF, in which meta learning is disabled. Especially for small privacy budgets, MetaMF yields a much more stable accuracy than NoMetaMF (*RQ2a*).

predictive capabilities well, even with decreasing privacy budgets. However, a privacy budget of $\approx 50\%$ seems to be a critical threshold. The ΔMAE@β only marginally increases for $\beta > 0.5$, but rapidly grows for $\beta \leq 0.5$ in the case of the Douban, Hetrec-MovieLens, and MovieLens 1M dataset. In other words, a user could afford to withhold $\leq 50\%$ of her data and still get well-suited recommendations. Additionally, the ΔMAE@β remains stable for the Ciao and Jester dataset. Similar observations can be made about the results of NoMetaMF in Fig. 2b. Again, the predictive capabilities remain stable for $\beta > 0.5$ in the case of Douban, Hetrec-MovieLens, and MovieLens 1M, but decrease tremendously for higher levels of privacy. Our side-by-side comparison of MetaMF and NoMetaMF in Fig. 2 suggests that both methods exhibit robust behavior for large privacy budgets (i.e., $\beta > 0.5$), but exhibit an increasing MAE for less data available (i.e., $\beta \leq 0.5$). However, we would like to highlight that the increase of the MAE is much worse for NoMetaMF than for MetaMF. Here, the ΔMAE@β indicates that the MAE for NoMetaMF increases much faster than the MAE for MetaMF for decreasing privacy budgets. This observation pinpoints the importance of meta learning and personalization in settings with a limited amount of data per user, i.e., a high privacy level. Thus, concerning *RQ2a*, we conclude that MetaMF is indeed more robust against decreasing privacy budgets than NoMetaMF, but yet, requires a sufficient amount of data per user.

Next, we compare MetaMF to NoMetaMF with respect to their ability for personalization and collaboration in the setting of decreasing privacy budgets. As explained in Sect. 3, we refer to Lin et al., which suggest that personalization leads to distinct weight embeddings and collaboration leads to clusters within the embedding space. In Fig. 3, we illustrate the weights of the first layer of the rating prediction models of MetaMF and NoMetaMF for the MovieLens 1M dataset for different privacy budgets (i.e., $\beta \in \{1.0, 0.5, 0.1\}$). Again, we applied t-SNE to reduce the dimensionality to two dimensions, followed by standardization to ease the visualization. In the case of MetaMF, we observe that it preserves the

Fig. 3. Weights of the first layer of the rating prediction models for the MovieLens 1M dataset. (a), (b), (c) depict MetaMF, whereas (d), (e), (f) depict NoMetaMF, in which meta learning is disabled. No well-defined clusters are visible for NoMetaMF, which indicates the inability to exploit collaborative information among users (*RQ2a*).

ability to generate different weights for each user for decreasing privacy budgets. Similarly, well-defined clusters can be seen, which indicates that MetaMF also preserves the ability to capture collaborative information among users. In contrast, our visualizations for NoMetaMF do not show well-defined clusters. This indicates that NoMetaMF loses the ability to exploit collaborative information among users. Due to limited space, we refrain from presenting the weights of the first layer of the rating prediction models for the other datasets. However, we observe that MetaMF outperforms NoMetaMF in preserving the collaboration ability for decreasing privacy budgets on the remaining four datasets, which is also in line with our previous results regarding *RQ2a*.

In the following, we elaborate on how the high degree of personalization in MetaMF impacts the recommendations of groups of users with different amounts of rating data. In a preliminary experiment, we measure the MAE on our three user groups *Low*, *Med*, and *High* on our five datasets in Table 2. Except for the Ciao dataset, our results provide evidence that *Low* is served with significantly worse recommendations than *High*. In other words, users with lots of ratings are advantaged over users with only a few ratings.

To detail the impact of decreasing privacy budgets on these user groups, we monitor the ΔMAE@β on *Low*, *Med*, and *High*. The results for our five datasets are presented in Fig. 4. Surprisingly, *Low* seems to be much more robust against small privacy budgets than *High*. Here, we refer to our observations about MetaMF's performance on the Ciao and Jester dataset in Fig. 2a. In contrast to the other datasets, Ciao and Jester comprise only a small average number of ratings per user, i.e., 38 (Ciao) and 56 (Jester), which means that they share a common property with our *Low* user group. Thus, we suspect a relationship

Table 2. MetaMF's MAE (mean absolute error) measurements for our three user groups on the five datasets. Here, we simulated a privacy budget of $\beta = 1.0$. According to a one-tailed t-Test, *Low* is significantly disadvantaged over *High*, indicated by *, i.e., $\alpha = 0.05$ and ****, i.e., $\alpha = 0.0001$ (*RQ2b*).

Dataset	Low	Med	High
Douban*	0.638	0.582	0.571
Hetrec-MovieLens****	0.790	0.603	0.581
MovieLens 1M****	0.770	0.706	0.673
Ciao	0.773	0.771	0.766
Jester****	1.135	0.855	0.811

between the robustness against decreasing privacy budgets and the amount of rating data per user. The most prominent examples of *Low* being more robust than *High* can be found in Figs. 4a, 4b and 4c. Here, the accuracy of MetaMF on *High* substantially decreases for small privacy budgets. On the one hand, MetaMF provides strongly personalized recommendations for users with lots of ratings, which results in a high accuracy for these users (i.e., *High*). On the other hand, this personalization leads to a serious reliance on the data, which has a negative impact on the performance in settings with small privacy budgets. Thus, concerning *RQ2b*, we conclude that users with lots of ratings receive better recommendations than other users if they can take advantage of their abundance

(a) Douban   (b) Hetrec-MovieLens   (c) MovieLens 1M
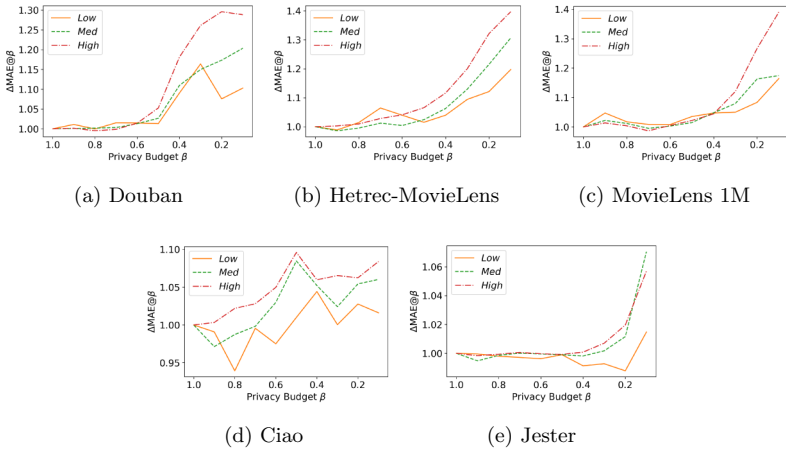
(d) Ciao   (e) Jester

Fig. 4. MetaMF's ΔMAE@β measurements for the (a) Douban, (b) Hetrec-MovieLens, (c) MovieLens 1M, (d) Ciao, and (e) Jester dataset for all three usergroups. Especially (a), (b), and (c) illustrate that *High* is sensitive to small privacy budgets. In contrast, *Low* can afford a high degree of privacy, since the accuracy of its recommendations only marginally decreases (*RQ2b*).

of data. In settings where a high level of privacy is required, i.e., a low privacy budget, and thus, users decide to hold back the majority of their data, users are advantaged who do not require as much personalization from the recommender system.

5   Conclusions and Future Work

In our study at hand, we conducted two lines of research. First, we reproduced results presented by Lin et al. in [16]. Besides, we introduced a fifth dataset, i.e., Jester, which, in contrast to the originally utilized datasets, has plenty of rating data per item. We found that all accuracy measurements are indeed reproducible ($RQ1a$). However, our reproduction of the t-SNE visualizations of the embeddings illustrated potential discrepancies between our and Lin et al.'s work ($RQ1b$). Second, we conducted privacy-focused studies. Here, we thoroughly investigated the meta learning component of MetaMF. We found that meta learning takes an important role in preserving the accuracy of the recommendations for decreasing privacy budgets ($RQ2a$). Furthermore, we evaluated MetaMF's performance with respect to decreasing privacy budgets on three user groups that differ in their amounts of rating data. Surprisingly, the accuracy of the recommendations for users with lots of ratings seems far more sensitive to small privacy budgets than for users with a limited amount of data ($RQ2b$).

Future Work. In our future work, we will research how to cope with incomplete user profiles in our datasets, as users may already have limited the amount of their rating data to satisfy their privacy constraints. Furthermore, we will develop methods that identify the ratings a user should share based on the characteristics of the data.

References

1. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.: The unfairness of popularity bias in recommendation. In: Workshop on Recommendation in Multistakeholder Environments in Conjunction with RecSys 2019 (2019)
2. Ammad-Ud-Din, M., et al.: Federated collaborative filtering for privacy-preserving personalized recommendation system. arXiv preprint arXiv:1901.09888 (2019)
3. Cantador, I., Brusilovsky, P., Kuflik, T.: Second international workshop on information heterogeneity and fusion in recommender systems. In: RecSys 2011 (2011)

4. Chen, C., Zhang, J., Tung, A.K., Kankanhalli, M., Chen, G.: Robust federated recommendation system. arXiv preprint arXiv:2006.08259 (2020)
5. Chen, F., Luo, M., Dong, Z., Li, Z., He, X.: Federated meta-learning with fast convergence and efficient communication. arXiv preprint arXiv:1802.07876 (2018)
6. Duriakova, E., et al.: PDMFRec: a decentralised matrix factorisation with tunable user-centric privacy. In: RecSys 2019 (2019)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML 2017 (2017)
8. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: AIS-TATS 2011 (2011)
9. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: a constant time collaborative filtering algorithm. Inf. Retrieval **4**(2), 133–151 (2001)
10. Guo, G., Zhang, J., Thalmann, D., Yorke-Smith, N.: ETAF: an extended trust antecedents framework for trust prediction. In: ASONAM 2014 (2014)
11. Ha, D., Dai, A., Le, Q.V.: Hypernetworks. In: ICLR 2016 (2016)
12. Hahnloser, R.H., Sarpeshkar, R., Mahowald, M.A., Douglas, R.J., Seung, H.S.: Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature **405**(6789), 947–951 (2000)
13. Harper, F.M., Konstan, J.A.: The movielens datasets: history and context. ACM Trans. Interact. Intell. Syst. (TIIS) **5**(4), 1–19 (2015)
14. Hu, L., Sun, A., Liu, Y.: Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In: SIGIR 2014 (2014)
15. Jiang, Y., Konečnỳ, J., Rush, K., Kannan, S.: Improving federated learning personalization via model agnostic meta learning. In: International Workshop on Federated Learning for User Privacy and Data Confidentiality in conjunction with NeurIPS 2019 (2019)
16. Lin, Y., et al.: Meta matrix factorization for federated rating predictions. In: SIGIR 2020 (2020)
17. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
18. Müllner, P., Kowald, D., Lex, E.: User Groups for Robustness of Meta Matrix Factorization Against Decreasing Privacy Budgets (2020). https://doi.org/10.5281/zenodo.4031011
19. Schedl, M., Bauer, C.: Distance-and rank-based music mainstreaminess measurement. In: UMAP 2017 (2017)
20. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NIPS 2017 (2017)
21. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Res. **30**(1), 79–82 (2005)