META-LEARNING AND TARGETED DIFFERENTIAL PRIVACY TO IMPROVE THE ACCURACY-PRIVACY TRADE-OFF IN RECOMMENDATIONS

Peter Müllner [1] Dominik Kowald [1,2] Markus Schedl [3,4] Elisabeth Lex [5]

[1]Know Center Research GmbH [2]University of Graz [3]Johannes Kepler University Linz [4]Linz Institute of Technology [5]Graz University of Technology

Problem and Idea

Recommender systems rely on user interaction data that often exhibits strong stereotypical patterns (e.g., gender or age). This data can reveal sensitive attributes via the generated recommendations.

Differential Privacy (DP) prohibits this by adding random noise to the data, but this severely degrades recommendation accuracy.

We solve this accuracy-privacy trade-off by applying DP only to strongly stereotypical, high-risk data (targeted DP) to reduce unnecessary noise, and additionally, we utilize meta-learning to improve the robustness to residual DP-noise.

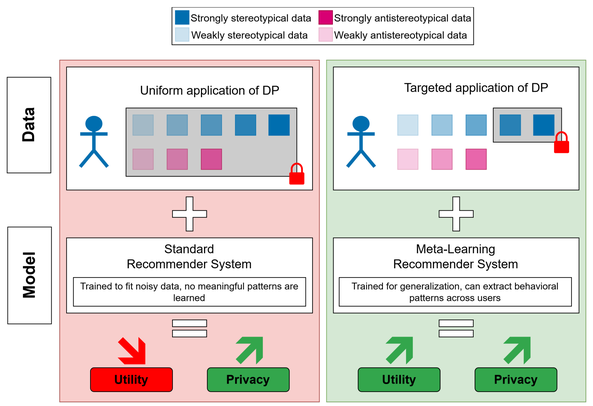

Figure 1. Our two-stage approach compared to traditional approaches.

Our Method

1. Define the Data Budget β: We introduce a data budget $\beta \in [0, 1]$ to control to how much user data DP is applied. β is the fraction of a user's profile (D_u) left untouched to preserve recommendation utility and $1 - \beta$ is the remaining fraction containing high-risk data that must be protected.

2. Select Top Stereotypical Items: The $k = \lceil (1 - \beta) \cdot |D_u| \rceil$ most stereotypical items (D_u^s) from a user's profile are selected. An item's risk level is measured via its Item-Group-Inclination [1] for a sensitive attribute a over its complement $\overline{a}$:

$$D_u^s = \arg \max_{r_{u,i} \in D_u}^{k} \mathbb{I}_a(u) \frac{IGI(i,a) - IGI(i,\overline{a})}{\max\{IGI(i,a), IGI(i,\overline{a})\}} \tag{1}$$

where $\mathbb{I}_a(u) = 1$ if user u possesses the sensitive attribute, and -1 otherwise.

3. Apply DP Mechanism m_{DP}: We apply a randomized coin-flip to D_u^s. With a probability determined by the privacy budget ϵ, the rating is replaced with a uniform random value ($\tilde{r}_{u,j}$) from an unvisited item ($j \sim \text{Uniform}(I \setminus I_u)$):

$$m_{DP}(r_{u,i}) = \begin{cases} r_{u,i} & \text{with prob. } \frac{e^\epsilon}{e^\epsilon + 1} \\ \tilde{r}_{u,j} & \text{with prob. } \frac{1}{e^\epsilon + 1} \end{cases} \tag{2}$$

4. Final Training Dataset $\tilde{D}_u$: The model trains on the merged dataset, combining the unprotected data with the DP-protected stereotypical data:

$$\tilde{D}_u = (D_u \setminus D_u^s) \cup \{m_{DP}(r_{u,i}) : r_{u,i} \in D_u^s\} \tag{3}$$

5. Train Meta-learning Recommendation Model: To address the remaining residual DP-noise, we use an existing meta-learning matrix factorization model [2] that can extract behavioral patterns across users despite noise.

Results

As more data is protected (lower data budget β), accuracy degradation increases. The most stereotypical items can be protected without causing an extra accuracy loss compared to random selection. Meta-Learning helps to reduce this degradation, which keeps recommendation accuracy high.

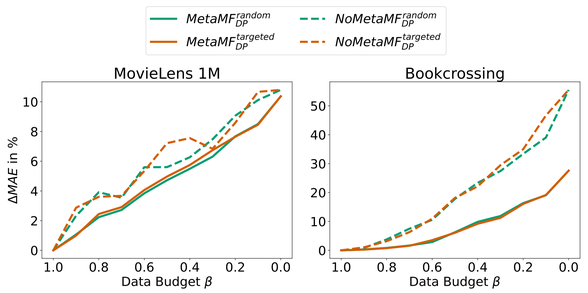

Figure 2. Recommendation accuracy for $\epsilon = 0.1$.

Especially for strict privacy regimes (i.e., $\epsilon = 0.1$), a targeted application of DP drops the attacker's inference accuracy to a minimum at $\beta = 0.3$. Here, protecting 70% of the data with DP drives profiles into a "neutral" state that masks stereotypical patterns significantly better than complete protection ($\beta = 0$).

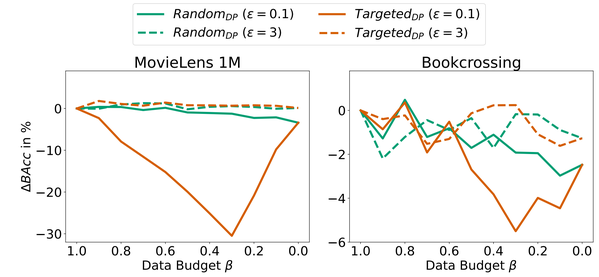

Figure 3. Empirical privacy risk for $\epsilon = 0.1$ and $\epsilon = 3$ (Neural Attacker).

Key-Takeaways

Source vs. Impact: In practice, not all data needs to be protected. We apply targeted DP to address the source of DP-noise, and use meta-learning to mitigate the impact of the remaining noise.

Neutrality Sweet Spot: The data budget acts as a powerful lever to reach "neutral" stereotypicality, which yields the best empirical privacy risk while keeping high-risk data secured under formal DP guarantees.

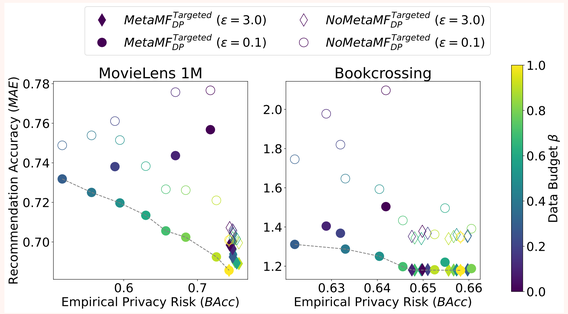

Figure 4. Accuracy-privacy trade-off.

References

[1] Gustavo Escobedo, Marta Moscati, Peter Muellner, Simone Kopeinik, Dominik Kowald, Elisabeth Lex, and Markus Schedl. Making alice appear like bob: A probabilistic preference obfuscation method for implicit feedback recommendation models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 349–365. Springer, 2024.

[2] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Dongxiao Yu, Jun Ma, Maarten de Rijke, and Xiuzhen Cheng. Meta matrix factorization for federated rating predictions. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 981–990, 2020.

pmuellner@know-center.at