An Analysis of the Impact of Differentially Private Data on Recommendation

Antonio Ferrara, Alberto Carlo Maria Mancino, Angela Di Fazio,
Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio

Ital-IA, Convegno Nazionale CINI sull'Intelligenza Artificiale
Pisa, 29–31 Maggio 2023

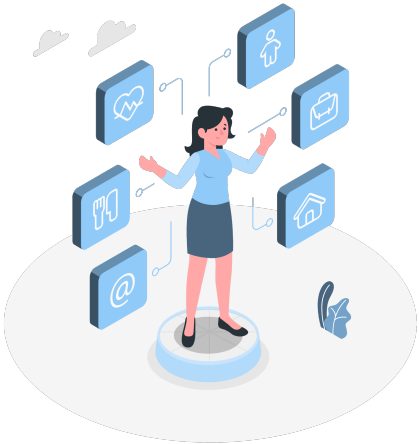
Motivations

Personalized Machine Learning

Machine Learning models are able to enhance the users' digital experience by providing **personalized services**

Unfortunately, they require a large number of users' personal data leading to several privacy risks

Nowadays **privacy protection** is a matter of great concern for users and governments



Recommender Systems

Outline

Recommender Systems are widely adopted information-filtering systems for mitigating the **information overload problem**

What to watch tomorrow?
What to buy now?
Where to go next?

Despite their **utility**, Recommender Systems cannot ignore the risks associated with the collection of **personal data**

Recommender Systems

Privacy matters

In order to learn users' behavioral patterns, Recommender Systems collect their **personal feedback**

They can be both explicit (5-star rating) and implicit (clickthrough data)

The more data collected, the better the ability to intercept users' tastes

Personalization-Privacy Paradox

- Users want accurate recommendation
- Users do not want to share their data

The exsisting trade-off between privacy and accuracy is a problem that must be addressed

Several works in the literature leverage the rigorous guarantees of **differential privacy** to build private models

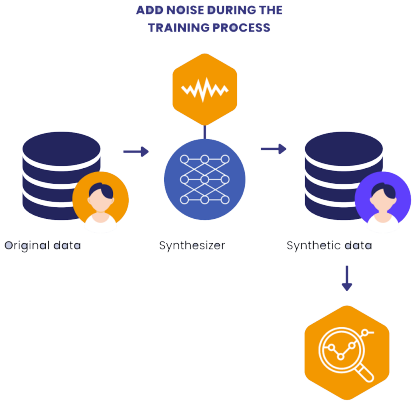
Differential Privacy

For Recommender Systems

Usually, Recommender Systems **integrate Differential Privacy** for noising model parameters or sharing noisy gradients

Few works apply differential privacy in the **data collection or release phases**

This aspect must not be underestimated since the majority of the models are trained in **offline settings**



Data Characteristics and Privacy

The intuition

The introduction of noise within the data inevitably **changes their structural characteristics** and impacts recommendation performance

As demonstrated by Adomavicius et al.[1], these characteristics, including sparsity and skewness, **play a fundamental role** in the recommendation system's performance, each with a different impact

Our intuition is that by analyzing the variation in dataset characteristics, it is possible to predict the impact of differential privacy on recommendation

[1] Adomavicius, Gediminas, and Jingjing Zhang. "Impact of data characteristics on recommender systems performance." ACM Transactions on Management Information Systems (TMIS) 3.1 (2012): 1-17.

Data Characteristics and Privacy

Our contribution

In our work, we carry out a preliminary **explanatory analysis** regarding the correlation between the characteristics of recommendation datasets and performance

We applied a simple **randomized-response-based** mechanism for privatizing implicit feedback data

We assess the impact of the original dataset characteristics and the chosen privacy strength on the **accuracy** and **popularity bias** on four different recommendation models

Privatizing User Data

Differential Privacy

Background

Differential Privacy represents a **formal mathematical definition of privacy**.
It is based on the principle that the output of computation should not allow inference about the presence of any particular individual

Differential Privacy

$\mathcal{M}$ is a randomized mechanism that takes as input a dataset d and returns a value in a space $\mathcal{O}$. $\mathcal{M}$ is said to satisfy DP if given any two adjacent datasets d_1 and d_2 and for any subset of possible outputs $\mathcal{S} \subseteq \mathcal{O}$, we have

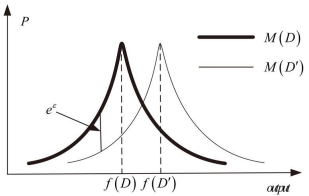
$$P(\mathcal{M}(D_1) \in \mathcal{S}) < e^{\epsilon} P(\mathcal{M}(D_2) \in \mathcal{S}) + \delta$$

Differential Privacy

Implications

As a result, regardless of whether or not a record is in the dataset, the ratio of the probabilities of the outcomes of $\mathcal{M}(D_1)$ and $\mathcal{M}(D_2)$ being in the same space S is bounded by e^{ϵ}

e^{ϵ} controls the **privacy budget** of a differentially private mechanism: **the smaller** the value of ϵ, **the stronger** the privacy guarantee for the mechanism.



Randomized Response

Background

Randomized Response is a mechanism that respondents to a survey can use to protect their privacy when asked about a **sensitive attribute**, e.g., «Did you visit Venice?».

Randomized Response

Let $x \in \{x_1, ..., x_r\}$ be the variable containing the true answer to a sensitive question. The randomized response privatizes the true answer reporting the value of a variable $\tilde{x}$ instead of x, based on a probability $p_{uv} = \Pr[\tilde{x} = v \mid x = u]$, for $u, v \in \{x_1, \ldots, x_r\}$, where u is the true value and v is the reported value.

Randomized Response

Binary Data

With Randomized Response, we properly perturbed the **original user rating matrix** by perturbing each binary implicit feedback independently, as it inherently represents a private answer to a different sensitive question

The probability of revealing the presence of feedback in the dataset depends on **the ratio of the probabilities of releasing the true value or not**

$$\max\left\{\frac{p_{00}}{p_{10}}, \frac{p_{11}}{p_{01}}\right\} \leq e^{\epsilon}$$

This approach guarantees users' privacy thanks to the connection between Randomized Response and Differential Privacy

Explanatory Analysis

Regression model

Which characteristics can influence the outcome?

We realize a **regression model** to analyze which **dataset characteristics** are more prone to influence the outcome of different recommendation models when the randomized response is applied to the original dataset. If a recommendation model performs μ on a dataset $\mathbf{X}$ for a specific metric, we argue that a perturbed dataset $\tilde{\mathbf{X}}$ with ϵ-differential privacy causes a **degradation of the performance** $\Delta\mu$, which we estimate as:

$$\Delta\mu(\mathbf{X}, \epsilon) = \theta_0 + \sum_{i \in \mathcal{C}} \theta_i g_i(\mathbf{X}) + \theta_\epsilon \epsilon,$$

where $\mathcal{C}$ is the set of the selected statistical characteristics, and the $g_i(\mathbf{X})$'s represent the same characteristics measured on the original dataset $\mathbf{X}$.

Dataset Characteristics

To build the regression model, we choose a subset of the recommendation dataset characteristics identified by Adomavicious et al.

Space Size, *Shape*, *User Ratings*, *Item Ratings*, *Item Gini*

Four of them intercept the information on the **dataset structure**, while the last one on the **feedback frequency** within the dataset

Experiments

Experimental Protocol

We generated **600 random sub-datasets** for each of the following three well-known recommendation datasets:

- MovieLens 1M
- Amazon Digital Music
- LibraryThing

Each of the **1800 generated sub-datasets** was privatized using the randomized response applied with three different ϵ values (3, 2, and 0.5).

Experimental Protocol (1/2)

The resulting **7200 datasets** have been used to train four recommendation models from four different categories:

- popularity-based (*MostPop*)
- distance-based (*ItemKNN*)
- autoencoder (*EASER*)
- graph (*RP3*β)

The explanatory model is trained on **28,800 experiments** to analyze the variation of accuracy and popularity bias with respect to the characteristics of the non-privatized datasets.

Experimental Protocol (2/2)

The adopted metrics are:

- **Precision@N**: it measures the fraction of N recommended items that is actually relevant to the user
- **Average Recommendation Popularity** (ARP@N): it calculates the average popularity of the items in the recommendation list

They have been calculated on the top-10 recommendation lists.
The regression variables have been normalized.

Findings

The impact on performance

Our explanatory model can explain the **68% of accuracy variation and 91% of popularity bias variation**. This demonstrates that the chosen characteristics effectively explain the performance variation

As expected, the **privacy budget plays a key role** in explaining the performance variation: a lower ϵ leads to an accuracy worsening and an increase in popularity bias

This means that a higher privacy guarantee leads to accuracy degradation and drives the **recommendation through popular items**

Findings

The impact on performance

A higher **Space Size**, leads to a lower impact on performance, meaning that *bigger datasets are less affected by noise*.

Item and User Ratings show that in datasets affected by popularity bias, the *bias is increased* and that *cold-start users are very sensitive to noise injection*.

Finally, the results for **Shape and Item Gini** are *largely statistically significant* and show the ability to explain both the accuracy and the popularity bias variations.

Conclusion

Conclusion

This study has extensively analyzed which dataset characteristics are more prone to influence the accuracy and popularity bias of different recommendation models when randomized response is applied to the original dataset.

This investigation has unveiled several insights and provided interesting suggestions to the researcher interested in protecting users' privacy but **further analysis are needed**.

Future studies could unveil these aspects and extend our analysis considering **non-linear explanatory models** and the effects of applying more sophisticated differential privacy techniques.

Thank you!

For further questions and information, don’t hesitate to contact us!

Alberto Carlo Maria Mancino alberto.mancino@poliba.it
Antonio Ferrara antonio.ferrara@poliba.it
Angela Di Fazio angela.difazio@poliba.it
Vito Walter Anelli vitowalter.anelli@poliba.it
Tommaso Di Noia tommaso.dinoia@poliba.it
Eugenio Di Sciascio eugenio.disciascio@poliba.it