What You Might Like To Read After Watching Interstellar (Technical Report)

Theorem 1. *Given an item t_i, the sensitivity of* `X-Sim` *is denoted by GS and the similarity between t_i and any arbitrary item t_j is denoted by* `X-Sim`(t_i, t_j)*. Then, the Private Replacement Selection (PRS) mechanism, which outputs t_j as the replacement with a probability proportional to $exp(\frac{\epsilon \cdot \texttt{X-Sim}(t_i,t_j)}{2 \cdot GS})$, provides ϵ-differential privacy.*

Proof. Consider two datasets D and D' which differ at one user, say u. We denote the `X-Sim` (t_i,t_j) in dataset D as $q(D, t_i, t_j)$ and $I(t_i)$ as the set of items in target domain with quantified `X-Sim` values. The global sensitivity (GS) is defined as $max_{D'}||q(D, t_i, t_j) - q(D', t_i, t_j)||_1$. Our PRS mechanism outputs an item t_j as a private replacement for t_i. Then, we have the following:

$$\frac{Pr[PRS(t_i, I(t_i), q(D, I(t_i))) = t_j]}{Pr[PRS(t_i, I(t_i), q(D', I(t_i))) = t_j]} = \frac{exp(\frac{\epsilon \cdot q(D,t_i,t_j)}{2 \cdot GS})}{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D,t_i,t_k)}{2 \cdot GS})} \div \frac{exp(\frac{\epsilon \cdot q(D',t_i,t_j)}{2 \cdot GS})}{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D',t_i,t_k)}{2 \cdot GS})}$$

$$= \underbrace{\frac{exp(\frac{\epsilon \cdot q(D,t_i,t_j)}{2 \cdot GS})}{exp(\frac{\epsilon \cdot q(D',t_i,t_j)}{2 \cdot GS})}}_{\text{P}} \cdot \underbrace{\frac{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D',t_i,t_k)}{2 \cdot GS})}{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D,t_i,t_k)}{2 \cdot GS})}}_{\text{Q}}$$

$$P = exp(\frac{\epsilon \cdot (q(D, t_i, t_j) - q(D', t_i, t_j))}{2 \cdot GS}) \leq exp(\frac{\epsilon \cdot GS}{2 \cdot GS}) = exp(\frac{\epsilon}{2})$$

$$Q = \frac{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D',t_i,t_k)}{2 \cdot GS})}{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D,t_i,t_k)}{2 \cdot GS})} \leq \frac{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot (q(D,t_i,t_k)+GS)}{2 \cdot GS})}{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D,t_i,t_k)}{2 \cdot GS})} = \frac{exp(\frac{\epsilon}{2}) \cdot \sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D,t_i,t_k)}{2 \cdot GS})}{\sum\limits_{t_k \in I(t_i)} exp(\frac{\epsilon \cdot q(D,t_i,t_k)}{2 \cdot GS})} = exp(\frac{\epsilon}{2})$$

Therefore, we have:

$$\frac{Pr[PRS(t_i, I(t_i), q(D, I(t_i))) = t_j]}{Pr[PRS(t_i, I(t_i), q(D', I(t_i))) = t_j]} \leq exp(\epsilon)$$

Hence, PRS provides ϵ-differential privacy. □

Theorem 2 (Recommendation-aware sensitivity)**.** *Given a score function $q : \mathcal{R} \rightarrow R$ and a dataset D, we define the recommendation-aware sensitivity corresponding to a score function $q_i(I, t_j)$ for a pair of items t_i and t_j as:*

$$RS(t_i, t_j) = max\{max_{u_x \in U_{ij}}(\frac{r_{t_{xi}} \times r_{t_{xj}}}{\parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel}), max_{u_x \in U_{ij}}(\frac{r_{t_i} \cdot r_{t_j}}{\parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} - \frac{r_{t_i} \cdot r_{t_j}}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel})\}$$

Proof. Now, we provide the proof of the recommendation-aware sensitivity. First, we have:

$$RS(t_i, t_j) = max \parallel s(t_i, t_j) - s'(t_i, t_j) \parallel_1$$

Next, we insert the similarity values for $s(t_i, t_j)$. A rating vector $r_{t_i} = [r_{t_{ai}}, ..., r_{t_{xi}}, r_{t_{yi}}]$ consists of all the ratings for an item t_i. Note that here a rating $r_{t_{xi}}$ denotes the result after subtracting the average rating of user x ($\bar{r_x}$) from the actual rating provide by x for an item i. Now, we have:

$$\begin{aligned} s(t_i, t_j) - s'(t_i, t_j) &= \frac{r_{t_i} \cdot r_{t_j}}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel} - \frac{r'_{t_i} \cdot r'_{t_j}}{\parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} \\ &= \frac{r_{t_i} \cdot r_{t_j} \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel - r'_{t_i} \cdot r'_{t_j} \times \parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} = \frac{P}{Q} \end{aligned}$$

Now, we assume that the profile of a user x, in D, is not present in D'. This user rated both t_i and t_j. Note that if this user rated one of these items or none, then the similarity value does not depend on the presence or absence of this user in the dataset. Hence, we have: $\parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel \leq \parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel$.

Now, we have P= $(r_{t_i} \cdot r_{t_j} \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel - r'_{t_i} \cdot r'_{t_j} \times \parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel)$ and Q=$(\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel)$. Hence, $Q \geq 0$ and depending on whether $P \geq 0$ or $P \leq 0$ we have two conditions which are as follows.

If $P \geq 0$, then we have:

$$\begin{aligned} \parallel s(t_i, t_j) - s'(t_i, t_j) \parallel_1 &= \frac{r_{t_i} \cdot r_{t_j} \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel - r'_{t_i} \cdot r'_{t_j} \times \parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} \\ &\leq \frac{(r_{t_i} \cdot r_{t_j} - r'_{t_i} \cdot r'_{t_j}) \times \parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} = \frac{(r_{t_i} \cdot r_{t_j} - r'_{t_i} \cdot r'_{t_j})}{\parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} \end{aligned}$$

If $P \leq 0$, then we have:

$$\begin{aligned} \parallel s(t_i, t_j) - s'(t_i, t_j) \parallel_1 &= \frac{r'_{t_i} \cdot r'_{t_j} \times \parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel - r_{t_i} \cdot r_{t_j} \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} \\ &= \frac{(r_{t_i} \cdot r_{t_j} - r_{t_{xi}} \times r_{t_{xj}}) \times \parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} - \frac{r_{t_i} \cdot r_{t_j} \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} \\ &= \frac{r_{t_i} \cdot r_{t_j} \times (\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel - \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel)}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} - \frac{r_{t_{xi}} \times r_{t_{xj}} \times \parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} \\ &\leq \frac{r_{t_i} \cdot r_{t_j} \times (\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel - \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel)}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel \times \parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} = \frac{r_{t_i} \cdot r_{t_j}}{\parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} - \frac{r_{t_i} \cdot r_{t_j}}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel} \end{aligned}$$

Hence, we have the recommendation-aware sensitivity as:

$$RS(t_i, t_j) = max\{max_{u_x \in U_{ij}}(\frac{r_{t_{xi}} \times r_{t_{xj}}}{\parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel}), max_{u_x \in U_{ij}}(\frac{r_{t_i} \cdot r_{t_j}}{\parallel r'_{t_i} \parallel \times \parallel r'_{t_j} \parallel} - \frac{r_{t_i} \cdot r_{t_j}}{\parallel r_{t_i} \parallel \times \parallel r_{t_j} \parallel})\}$$

□

Theorem 3. *Given an item t_i, we denote its k neighbors by $N_k(t_i)$, the maximal length of all the rating vector pairs by $|v|$, the minimal similarity among the items in $N_k(t_i)$ by $Sim_k(t_i)$ and the maximal recommendation-aware sensitivity between t_i and other items by RS. Then, for a small constant $0 < \rho < 1$, the similarity of all the items in $N_k(t_i)$ are larger than $(Sim_k(t_i) - w)$ with a probability at least $1 - \rho$, where $w = min(Sim_k(t_i), \frac{4k \times RS}{\epsilon'} \times ln\frac{k \times (|v| - k)}{\rho})$.*

Proof. First, we recall that the probability of selecting an item is allocated by:

$$P = \frac{exp(\frac{\epsilon' \times \widehat{Sim}(t_i,t_j)}{4k \times RS(t_i,t_j)})}{\sum\limits_{l \in C_1} exp(\frac{\epsilon' \times \widehat{Sim}(t_i,t_l)}{4k \times RS(t_i,t_l)}) + \sum\limits_{l \in C_0} exp(\frac{\epsilon' \times \widehat{Sim}(t_i,t_l)}{4k \times RS(t_i,t_l)})}$$

where $\widehat{Sim(t_i,t_j)} = max(Sim(t_i,t_j), Sim_k(t_i) - w)$, $C_0 = [t_j | s(t_i,t_j) < Sim_k(t_i) - w, t_j \in I]$ and $C_1 = [t_j | s(t_i,t_j) \geq Sim_k(t_i) - w, t_j \in I]$.

We begin our proof by computing the probability of selecting a neighbor with a similarity less than $Sim_k(t_i) - w$ in each round of sampling. Given that a neighbor with similarity of $Sim_k(t_i) - w$ is still waiting for selection, the probability (p) of selecting a neighbor with similarity less than $Sim_k(t_i) - w$ is:

$$p = \frac{exp(\frac{\epsilon' \times (Sim_k(t_i) - w)}{4k \times RS})}{\sum\limits_{l \in C_1} exp(\frac{\epsilon' \times \widehat{Sim}(t_i,t_l)}{4k \times RS}) + \sum\limits_{l \in C_0} exp(\frac{\epsilon' \times \widehat{Sim}(t_i,t_l)}{4k \times RS})} \leq \frac{exp(\frac{\epsilon' \times (Sim_k(t_i) - w)}{4k \times RS})}{exp(\frac{\epsilon' \times Sim_k(t_i)}{4k \times RS})} = exp(\frac{-\epsilon' w}{4k \times RS})$$

The inequality is due to the fact that the k^{th} neighbor, with a similarity $Sim_k(t_i)$, is still present in $C_0 \cup C_1$ during the selection.

Since there are at most $|v|$ neighbors whose similarities are less than $Sim_k(t_i) - w$, the probability of choosing a neighbor with similarity s less than $Sim_k(t_i) - w$ is at most $(|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS})$. Furthermore, the probability of choosing any neighbor with similarity less than $Sim_k(t_i) - w$ leads to the following inequality after k sampling rounds:

$$P_{s < Sim_k(t_i) - w} \leq (|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS}) \implies (1 - P_{s < Sim_k(t_i) - w})^k \geq [1 - (|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS})]^k$$

Here, $1 - P_{s < Sim_k(t_i) - w}$ denotes the probability of selecting a neighbor with similarity at least $Sim_k(t_i) - w$. By Bernoulli's inequality, we can obtain a subsequent inquality as follows.

$$(1 - P_{s < Sim_k(t_i) - w})^k \geq 1 - k \times (|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS}) \tag{1}$$

If we assume that $1 - k \times (|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS}) \geq 1 - \rho$, then the inequality leads to the following inference.

$$\begin{aligned} \rho \geq k \times (|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS}) &\implies \frac{\rho}{k \times (|v| - k)} \geq exp(\frac{-\epsilon' w}{4k \times RS}) \\ &\implies \ln(\frac{\rho}{k \times (|v| - k)}) \geq \frac{-\epsilon' w}{4k \times RS} \\ &\implies w \geq \frac{4k \times RS}{\epsilon'} \times \ln \frac{k \times (|v| - k)}{\rho} \end{aligned}$$

Hence, we observe that with a probability of at least $1 - \rho$, the similarity of all items in $N_k(t_i)$ are larger than $Sim_k(t_i) - w$. In practice, we have to ensure that the truncated similarities satisfies

the range of values i.e. $Sim_k(t_i) - w \geq 0$ for cosine or $Sim_k(t_i) - w \geq -1$ for pearson correlation, so $w = min(Sim_k(t_i), \frac{4k \times RS}{\epsilon'} \times \ln \frac{k \times (|v|-k)}{\rho})$.

□

Theorem 4. *Given an item t_i, for a small constant $0 < \rho < 1$, all items with similarities greater than $(Sim_k + w)$ are present in $N_k(t_i)$ with a probability at least $1-\rho$ where $w = min(Sim_k(t_i), \frac{4k \times RS}{\epsilon'} \times ln \frac{k \times (|v|-k)}{\rho})$.*

Proof. Let A denote the event that a neighbor with a similarity greater than $Sim_k(t_i) + w$ has not been selected in $N_k(t_i)$ and B denote the event of selecting a neighbor with similarity less than $Sim_k(t_i)$. Then, we have the following inequality.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \leq \frac{P(B)}{P(A)} = \frac{exp(\frac{\epsilon' \times Sim_k(t_i)}{4k \times RS})}{exp(\frac{\epsilon' \times (Sim_k(t_i)+w)}{4k \times RS})} = exp(\frac{-\epsilon' w}{4k \times RS})$$

Therefore, the probability $P_{s<Sim_k(t_i)}$ of selecting any neighbor with similarity less than $Sim_k(t_i)$ given the condition of the event A is as follows.

$$P_{s<Sim_k(t_i)} \leq (|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS})$$

In any of the k rounds of sampling, we can have the following inequality.

$$(1 - P_{s<Sim_k(t_i)})^k \geq [1 - (|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS})]^k \quad \geq 1 - k \times (|v| - k) \times exp(\frac{-\epsilon' w}{4k \times RS}) \geq 1 - \rho$$

As a result, we again have the inequality: $w \geq \frac{4k \times RS}{\epsilon'} \times \ln \frac{k \times (|v|-k)}{\rho}$.

Thus, similar to the proof of the previous theorem, when $w = min(Sim_k(t_i), \frac{4k \times RS}{\epsilon'} \times ln \frac{k \times (|v|-k)}{\rho})$, all the neighbors of t_i whose similarity are greater than $Sim_k(t_i) + w$ are present in $N_k(t_i)$ with a probability at least $1 - \rho$.

□