An Additional Definition

Jie Zhang, Yifan Dong, Li Yin, Zhiwu Li

Differential privacy is a privacy-preserving method with a rigorous mathematical definition, which offers a mechanism (or a function) that publishes aggregate information about a statistical database, where the private information in it is protected or restricted. In other words, if a dataset is considered as an input of a differential privacy mechanism, the addition or deletion of any one record (or element) in it does not affect the query result, i.e., an intruder cannot capture the private information with the slight modification of the dataset.

Definition 1 (Differential Privacy) Let ϵ be a positive real number and $\mathcal{K}_o$ be a randomized mechanism (function) that takes a dataset as input. Let $Im(\mathcal{K}_o)$ denote the image of $\mathcal{K}_o$. The mechanism $\mathcal{K}_o$ provides ϵ-differential privacy if for any two datasets $\mathcal{O}$ and $\mathcal{O}'$ that differ on a single element, for all $\mathcal{S} \subseteq Im(\mathcal{K}_o)$, it holds

$$P_r\left[\mathcal{K}_o(\mathcal{O}) \in \mathcal{S}\right] \leq \exp(\epsilon) P_r[\mathcal{K}_o(\mathcal{O}') \in \mathcal{S}],$$

where the value of $\mathcal{K}_o$ at a dataset $\mathcal{O}$ or $\mathcal{O}'$ is contained in the sample space, i.e., $Im(\mathcal{K}_o)$, with a probability decided by the randomness used in the mechanism, and $P_r[\mathcal{K}_o(\mathcal{O}) \in \mathcal{S}]$ is the probability of $\mathcal{K}_o(\mathcal{O}) \in \mathcal{S}$ representing that the output of $\mathcal{K}_o$ at $\mathcal{O}$ belongs to $\mathcal{S}$. $\diamondsuit$

In Definition 1, the value ϵ evaluates the performance of differential privacy. Namely, a smaller value of ϵ implies a finer difference between the probabilities of $\mathcal{K}_o(\mathcal{O}) \in \mathcal{S}$ and $\mathcal{K}_o(\mathcal{O}') \in \mathcal{S}$, i.e., the intruder is less likely to distinguish the two datasets. On the contrary, a larger ϵ means a lower degree of users' private information protection.

Remark 1 *By Definition 1, the two datasets $\mathcal{O}$ and $\mathcal{O}'$ are required to be different on a single element. However, this work introduces differential privacy in the field of DESs modeled by DFAs $G_d = (Q_d, \Sigma, f_d, q_0)$, and defines $\mathcal{K} : \mathcal{Q}_d^l \to \mathcal{Q}_d^l$ as the mechanism specifically defined on a set of state sequences (the formal definition of differential privacy in the framework of DFAs, called state sequence differential privacy is presented in Section 4). In the modified definition of differential privacy catering for a DES, the input and output of $\mathcal{K}$ are two state sequences with the same length, whose similarity is measured by the Hamming distance (defined in Section 4) between them.*

Currently, three basic mechanisms are widely used to ensure differential privacy depending on the forms of the query that is a function mapping a dataset $\mathcal{O}$ to an abstract range. In particular, the Laplace and Gaussian mechanisms are suitable for a numeric query, i.e., the output of the query belongs to a real number (or a real vector), while the exponential mechanism is applicable for a non-numeric query, whose output is a non-numerical element. In our work, we mainly touch upon the problem of differential privacy for state sequences, where both the domain and codomain of the random mechanism $\mathcal{K}$ are $\mathcal{Q}_d^l$. Namely, the output of $\mathcal{K}$ is non-numeric.