차분 프라이버시의 노이즈 메커니즘에 대한 분석

우타리예바 아쎔*, 신진명*, 최윤호**

*부산대학교 (대학원생), **부산대학교 (교수)

A Survey on Noise Mechanisms of Differential Privacy

Utaliyeva Assem*, Jinmyeong Shin*, Yoon-Ho Choi**

*Pusan National University(Graduate student),
**Pusan National University(Professor)

Abstract

Since the threat to data privacy is increasing, preserving data privacy is an important issue of data publishing and mining tasks. Recently, Differential Privacy emerged as a state-of-art concept that provides strong mathematical guarantees. However, understanding Differential Privacy is not an easy task. In this paper, to help our readers understand Differential Privacy, we introduce concept of Differential Privacy and major noise mechanisms to achieve Differential Privacy.

I.Introduction

Since the threat to data privacy is increasing, preserving data privacy is an important issue of data publishing and mining tasks. k-anonymity, l-diversity, and t-closeness are widely used to preserve data privacy. However, according to many research[3], such privacy models are not enough to guarantee data privacy.

Different from such conventional models, Differential Privacy(DP), which is introduced by C. Dwork, is resistant to existing data privacy attacks and guarantees data privacy mathematically. The intuitive concept of DP is to add randomness to all data which is published to third party. As shown in Fig. 1, Answers from database D1 and D2, which are different, but satisfy DP, are similar in bound of exp(ε)+δ. This characteristic makes it impossible to recognize the difference between two databases.

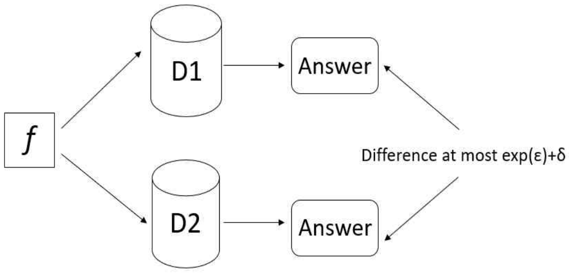

Fig. 1 An Conceptual Overview of Differential Privacy

Because of such an advantage, DP is widely studied and applied in many fields. However, since the way of achieving DP is a mathematically hard problem, it is not an easy task to understand the detailed concept of DP. To handle such problem, we introduce a detailed concept of DP and summarize noise mechanisms to achieve DP.

The paper is organized as follows. First, we explain the concept of DP in section 2. Then, representative noise mechanisms are described in section 3. In section 4, we compare the pros and cons of each noise mechanism. Then, we summarize and conclude the paper in section 5.

II. Differential Privacy

In this section, we describe key concepts of DP, which are DP's definition and sensitivity.

Definition 1 (ε, δ- differential privacy) A randomized algorithm K gives ε, δ- DP if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(K)$,

$$\frac{\Pr[K(D_1) \in S]}{\Pr[K(D_2) \in S]} \leq \exp(\epsilon) + \delta \qquad (1)$$

Equation 1 means that the presence or absence of a user in the dataset can cause at most exp(ε) + δ change. Here, ε is a privacy parameter, and δ is a relaxation parameter. A strict version is called ε-DP when δ is 0.

Definition 2 Sensitivity of a function $f: D \rightarrow R^w, w \in N^+$ is the maximum difference that absence of one individual in the data set can change.

$$\triangle f = \max \| f(D_1) - f(D_2) \|_1 \qquad (2)$$

$$\triangle f^2 = \max \sqrt{\| f(D_1) - f(D_2) \|^2} \qquad (3)$$

The equation 2 is definition L1-sensitivity and the equation 3 is definition of L2-sensitivity.

III. Noise Mechanisms

3.1 Laplace Mechanism

Laplace Mechanism is the most common mechanism to satisfy ε-DP. It perturbs the output by adding random noise to the true result of the query. The noise can be drown using the probability density function(PDF) in equation 4.

$$Lap(x|b) = \frac{1}{2b} exp(-\frac{|x|}{b}) \qquad (4)$$

The amount of noise is calibrated according to the privacy parameter ε and the sensitivity of the query $\triangle f$. The result of query is same as equation 5.

$$\tilde{f}(x) = f(x) + Lap(\frac{\triangle f}{\epsilon}) \qquad (5)$$

3.2 Gaussian Mechanism

Gaussian Mechanism is another output perturbation mechanism that uses the noise from Gaussian Normal Distribution. The PDF of Gaussian Noise is described in equation 6.

$$Gaussian(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad (6)$$

Random noise can be calculated by substituting values of mean μ = 0 and $\sigma^2 = (2\ln(1.25/\delta) \times \triangle f^2)/\epsilon^2$. Therefore, the result of query is same as equation 7.

$$\tilde{f}(x) = f(x) + Gaussian(0, \sigma) \qquad (7)$$

3.3 Exponential Mechanism

The main idea behind the Exponential Mechanism is the utility function $u: N^{|D|} \times R \rightarrow R$. The utility function $u(D, r)$ represents how good output r is for database D. As shown in equation 8, the mechanism selects and outputs an element $r \in R$ with probability.

$$\Pr(r) \propto \exp(\frac{\epsilon u(D,r)}{2\triangle u}) \qquad (8)$$

In this equation, $\triangle u$ is L1-sensitivity of utility function u.

Mechanism	Laplace	Gaussian	Exponential
Differential Privacy	ε-DP	ε,δ-DP	ε-DP
Data Type	Numeric	Numeric	Categorical
Mechanism Type	Output Perturbation	Output Perturbation	Scoring function
Additive Noise	$Lap(\frac{\triangle f}{\epsilon})$	$Gaussian(0,\sigma)$	-
Sensitivity	L1 norm	L2 norm	-

Table 1. Comparison of Differentially Private Mechanisms

IV.Comparison of Mechanisms

Mechanisms to achieve DP can be categorized into output perturbation and scoring function. The output perturbation type adds additive noise to the query output. Differently, the scoring function type returns the stochastic best element which satisfies the query.

One major difference between two output perturbation mechanisms, which are Laplace and Gaussian, is the sensitivity. Since two mechanisms are using different norms as sensitivities, the Laplace mechanism shows better data utility when the sensitivity is low. On the contrary, the Gaussian mechanism shows better data utility when the sensitivity is high.

For instance, if the user will affect just one statistics the both L1 and L2 sensitivities of the query will be equal to 1. In this case, the Laplace mechanism adds $Lap(1/\epsilon)$ amount of noise. This is much less than the noise generated by the Gaussian mechanism. However, when one user affects more than one statistic, the value of L2-sensitivity becomes less than L1-sensitivity. In the case of the 50 counting queries, the L1-sensitivity of a query is 50, whereas the L2-sensitivity of a query is ~7.07. Thus, the Gaussian mechanism is more appropriate to use in cases with multiple complex statistics in order to prevent high amounts of additive noise and degradation of the data utility.

Since output perturbation types are suitable only for numeric queries, the Exponential mechanism, which is the scoring function type, is designed to handle categorical data. Exponential mechanism does not add any noise directly to the answer but returns the best element that satisfies the condition. To satisfy DP, it sometimes returns the element with not the highest score. The mechanism provides higher utility since no additive noise was used.

V. Conclusion

In this paper, we reviewed the basic noise generating mechanisms of DP and their features. All previous work was focused on designing tailored mechanisms for specific data analysis. Since the application of DP is expanding to different areas, the regular noise generating mechanisms are not practical enough.

Acknowledgement

This research was supported by the MSI T, Korea, under the ITRC support program(II TP-2020-0-01797) and BK21 Four, Korean S outheast Center for the 4th Industrial Revol ution Leader Education.

[References]

[1] C.Dwork, A.Roth, The Algorithmic Foundations of Differential Privacy, 2014

[2] F.McSherry, K.Talwar, Mechanism Design via Differential Privacy, 2007

[3] C.Clifton, T.Tassa, On Syntactic Anonymity and Differential Privacy, 2013