J Inf Process Syst, Vol.20, No.4, pp.524~534, August 2024
https://doi.org/10.3745/JIPS.04.0316

ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Privacy-Constrained Relational Data Perturbation: An Empirical Evaluation

Deokyeon Jang, Minsoo Kim, and Yon Dohn Chung*

Abstract

The release of relational data containing personal sensitive information poses a significant risk of privacy breaches. To preserve privacy while publishing such data, it is important to implement techniques that ensure protection of sensitive information. One popular technique used for this purpose is data perturbation, which is popularly used for privacy-preserving data release due to its simplicity and efficiency. However, the data perturbation has some limitations that prevent its practical application. As such, it is necessary to propose alternative solutions to overcome these limitations. In this study, we propose a novel approach to preserve privacy in the release of relational data containing personal sensitive information. This approach addresses an intuitive, syntactic privacy criterion for data perturbation and two perturbation methods for relational data release. Through experiments with synthetic and real data, we evaluate the performance of our methods.

Keywords

Anonymization, Data Perturbation, Data Privacy, Personal Data Protection

1. Introduction

Data privacy over statistical databases is important, because databases contain personal sensitive information that should not be revealed when analyzing and publishing data [1-5]. There have been various data privacy studies: syntactic privacy methods [6-8] (such as k-anonymity, l-diversity, t-closeness, etc) and semantic privacy methods [9-11] (such as ε-differential privacy, (ε, δ)-differential privacy, μ-Gaussian differential privacy, etc).

The syntactic privacy research focuses on preventing privacy breaches directly from data, whereas the semantic privacy does from programs (or algorithms) processing the data. That is, the former is used for publishing (or releasing) data in an anonymized manner, and the latter is used for analyzing (or deep-learning) data in a privacy-preserving manner (not releasing data). Although there are some studies [12-15] generating and publishing synthetic data using the notion of differential privacy, they assume the usage of the released data (e.g., clustering, regression, and so on) is known a priori.

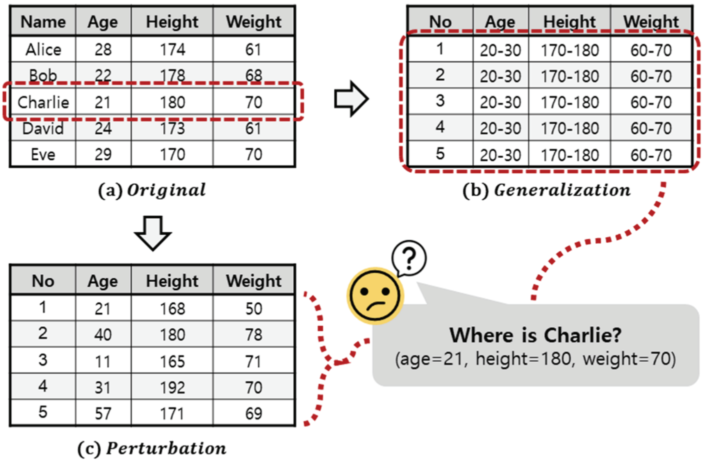
In the paper, we focus on the privacy-preserving relational data publication (a.k.a. de-identification) that modifies the source relation such that no individual record can be identified from the released relation. Although the semantic privacy (mostly the differential privacy) has become the major trend

Manuscript received September 26, 2023; first revision November 6, 2023; accepted November 12, 2023.

* **Corresponding Author:** Yon Dohn Chung (ydchung@korea.ac.kr)
Dept. of Computer Science & Engineering, Korea University, Seoul, Korea (deokyeonjang@korea.ac.kr, msdb@korea.ac.kr, ydchung@korea.ac.kr)

among academic researchers in recent years, the syntactic privacy issue (i.e., data release after anonymization/de-identification) has still been recognized its importance in industrial fields. Especially, in Korea, the data exchange and combination after pseudonymization/anonymization has been proliferated in recent years and to top it off is encouraged by the government [16,17].

For privacy-preserving data publication, two kinds of approaches have been explored in the past: (1) generalization-based approach [6-8] and (2) perturbation-based approach [16,18-20]. The first is to modify data records such that multiple data records have the same values for the QI (quasi-identifier) attributes [6]. For example, the k-anonymity model [6,7] requires every data record has at least "$k-1$" other data records with the same contents. Fig. 1 shows an example. Suppose there are five data records with the same height, weight and age values in the published relation. The individuals for these five records are not distinguishable (Fig. 1(b)), that is, de-identified. On the other hand, the second is to modify the data such that the published values are not the same to the original ones (Fig. 1(c)). Usually, for the perturbation (i.e., changing the data values), randomly generated noises are used. In spite of very different policies they adopt, both approaches achieve the same goal—anonymization.



Name	Age	Height	Weight
Alice	28	174	61
Bob	22	178	68
Charlie	21	180	70
David	24	173	61
Eve	29	170	70

(a) *Original*

No	Age	Height	Weight
1	20-30	170-180	60-70
2	20-30	170-180	60-70
3	20-30	170-180	60-70
4	20-30	170-180	60-70
5	20-30	170-180	60-70

(b) *Generalization*

No	Age	Height	Weight
1	21	168	50
2	40	180	78
3	11	165	71
4	31	192	70
5	57	171	69

(c) *Perturbation*

Fig. 1. Two anonymization approaches: generalization vs. perturbation.

Comparing the two approaches, it is possible to enforce the degree of anonymity in a quantitative manner in the generalization-based approach. The larger k (and also l) values denote the stronger anonymity [6,7]. However, there are situations such as when the data exhibits high-dimensionality, where achieving k-anonymity via generalization may not be feasible [21]. In particular, the dataset consists of numerous attributes that can be quasi-identifiers pose challenges in attaining de-identification without significant amount of data loss. In such cases, the perturbation-based approach may be preferred as it can provide privacy protection while accommodating the limitations posed by the data's high dimensionality. On the other hand, unlike the generalization-based privacy models such as k-anonymity [6] and l-diversity [7], the perturbation-based approach has a problem of lacking in the way of quantitatively controlling the leakage of privacy. In practical applications, the absence of a quantitative privacy measure is a significant obstacle to adoption; therefore, it is important to define privacy quantification [22].

In this respect, this paper proposes a privacy-constrained data perturbation method, namely ρ-RDP; especially we are focusing on relational data perturbation via additive noises. Fig. 2 shows the overview of our proposed method. Given a relation, we perturb the data via adding random noises. Here, we measure the amount of possible privacy leakage from the perturbed data and guarantee it satisfies the specified criteria, the ρ-safeness measure, that will be explained in Section 2. Then the output data can be safely utilized by the public.

The rest of the paper is organized as follows. Firstly, in Section 2, we define a novel syntactic privacy measure ρ for data perturbation. As far as we know, there has been no research defining a syntactic privacy metric for data perturbation. And we propose two data perturbation methods satisfying the given privacy constraint in Section 3. Lastly in Section 4, we evaluate the proposed methods with various experiments in comparison with the generalization-based anonymization method.

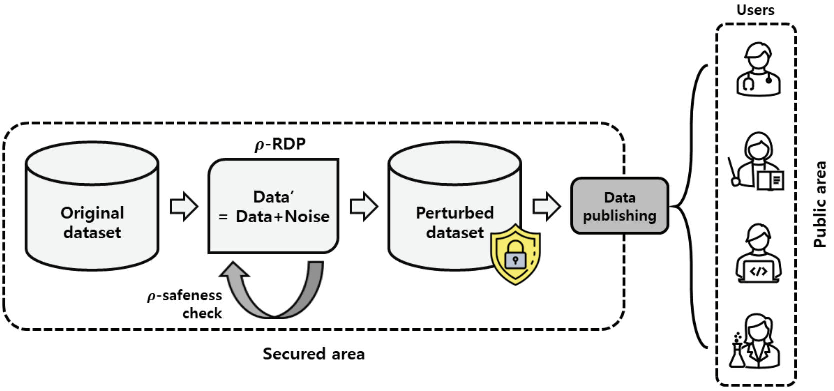


Fig. 2. The overview of our proposed method ρ-RDP.

2. Problem Definition

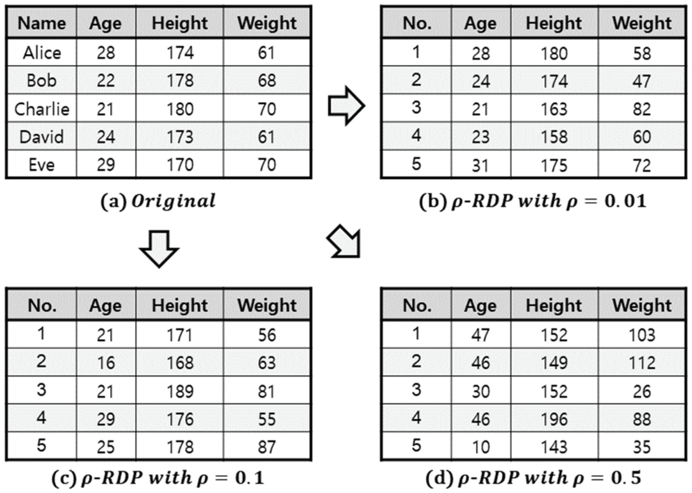
In this section, we formally define the problem of RDP that we tackle in the paper. We assume a relation is given where all attributes are in continuous and non-negative numeric domains in range [0, 1]. In practice, any numeric values can be used without loss of generality with appropriate scaling and transforming. Categorical attributes can also be used if relevant scoring functions and distance measures are specified. We assume the relation has no identifying attributes.

Definition 1 (RDP): Let R be a relation consisting of N records, where each record r_i is defined as $r_i = [r_{i1}, r_{i2}, \dots, r_{ij}, \dots, r_{iM}]$, where $i = 1$ to N, $j = 1$ to M and $0 \le r_{ij} \le 1$. The perturbation of r_i into r'_i means the addition of a random noise in range $[-1, 1]$ to each attribute's value like: $r'_{ij} = r_{ij} + X_{ij}$. Then, the perturbed relation, denoted as R', is the set of N records $r'_i = [r'_{i1}, r'_{i2}, \dots, r'_{ij}, \dots, r'_{iM}]$, where X_{ij} follows a random distribution. □

RDP just randomizes attribute values of original data records into those of result relation in a one-to-one way. Due to the randomness, the privacy leakage is restricted in probabilistic ways [1,2,4,23]. However, the deficiency of privacy quantification [24] makes it challenging for practitioners in the field to readily adopt. In this regard, we propose an intuitive, syntactic privacy criterion through which the amount of privacy leakage from r'_i is measured solely based on its original record r_i.

Definition 2 (ρ-safeness): A perturbed record r_i' is called "ρ-safe" compared with its original record r_i if the dissimilarity between r_i and r_i' is greater than ρ $(0 \le \rho \le 1)$, where $dissimilarity(r_i, r_i') = \sum_{j=1}^{M} |r_{ij} - r_{ij}'|/M$. □

The dissimilarity measure represents the ratio of distance between two records r_i and r_i' to the size of multidimensional space. We use the *Manhattan* (L_1) distance for the dissimilarity between two records assuming the set of M attribute values as a point in an M dimensional, normalized space. The privacy safeness between $r(0.3)$ and $r'(0.7)$ is 0.4. The record $r'(0.1, 0.4)$ is 0.45 privacy-safe compared with its original record $r(0.7, 0.1)$ (Interestingly, the distance concept has been used as the data utility measure, not privacy one, in the literature [6,7,11,15]. This is because they control the risk of re-identification by the size (i.e., "k" or "l") of so-called equivalence class). Fig. 3 shows some examples of ρ-safe perturbation results, where ρ values are 0.01, 0.1 and 0.5.



Name	Age	Height	Weight
Alice	28	174	61
Bob	22	178	68
Charlie	21	180	70
David	24	173	61
Eve	29	170	70

(a) $\boldsymbol{Original}$

No.	Age	Height	Weight
1	28	180	58
2	24	174	47
3	21	163	82
4	23	158	60
5	31	175	72

(b) $\boldsymbol{\rho\text{-}RDP\ with\ \rho = 0.01}$

No.	Age	Height	Weight
1	21	171	56
2	16	168	63
3	21	189	81
4	29	176	55
5	25	178	87

(c) $\boldsymbol{\rho\text{-}RDP\ with\ \rho = 0.1}$

No.	Age	Height	Weight
1	47	152	103
2	46	149	112
3	30	152	26
4	46	196	88
5	10	143	35

(d) $\boldsymbol{\rho\text{-}RDP\ with\ \rho = 0.5}$

Fig. 3. Examples of ρ-safe perturbation results.

Definition 3 (ρ-RDP): Given a relation R with N records and M attributes, the problem of privacy-constrained RDP with privacy safeness ρ is to generate a relation R' from R such that R' contains N data records r_i' and M attributes' values r_{ij}' for each record; $r_{ij}' = r_{ij} + X$, where X is a noise in range $[-1, 1]$ and $dissimilarity(r_i, r_i') > \rho$. Additionally, the noise X is generated by the function f, which produces M noises for each record based on specific distribution.

ρ-RDP randomizes the attribute values of original data records into those of result relation in a one-to-one way such that all records are "ρ-safe." Unlike the basic RDP, the amount of privacy protection (i.e., privacy riskiness) is quantitatively guaranteed via the ρ-safeness measure.

3. Privacy-Constrained Data Perturbation Methods

Algorithm 1 shows the algorithm of the ρ-RDP framework satisfying our proposed privacy constraint ρ-safeness. The algorithm adds noises to each (Line 5) and checks the privacy constraint (Lines 7–8; if necessary, the domain integrity check and relevant conversion can be added here according to the target applications). If the perturbed value does not satisfy the privacy constraint, we retry the perturbation. For the noise generation, we use the value distortion methods in [24] and consider the following two approaches:

1. ***Uniform noise***: $|X| \sim U\ (0,1)$
2. ***Laplace noise***: $X \sim Lap\ (\rho, 0.1)$

Algorithm 1. ρ-RDP algorithm

Input: the original relation R, the privacy parameter ρ and noise generation function f
Output: the perturbed relation R'

```
for $r_i$ in $R$ do
again:
    $X$ is generated by function $f$
    for $r_{ij}$ in $r_i$ do
        $r'_{ij} = r_{ij} + X$;
    end
    if $dissimilarity(r_i, r'_i) \leq \rho$ then
        goto again;   // privacy constraint violation
end
```

The former is a very straightforward approach for generating additive noises. A value in range $[0, 1]$ is randomly chosen, and added/subtracted to the original attribute value r_{ij}. However, although we need noises for providing ρ-safeness, too much noise is not desirable, since they degrade the data utility of the released relation. Under the restriction of satisfying the required ρ-safeness (i.e., $\sum_{j=1}^{M} \left| r_{ij} - r'_{ij} \right| > \rho \cdot M$), we should minimize the error (i.e., difference between the original and perturbed values). In order to provide adequate amount of noise, we devise the second method: generating noises from the Laplace distribution of mean ρ (we use the scale parameter "0.1" for simplicity.) and added or subtracted to the original attribute value with a half chance. This is based on the observation that the *Chebyshev* (L_∞) distance is equal to $Max|r_{ij} - r'_{ij}|$. That is, in order to guarantee the required privacy constraint with less errors, we try to add random noises whose absolute values are close to ρ as much as possible and distribute them evenly over M attributes.

4. Performance Evaluation

In this section, we carry out experiments for performance evaluation and describe the results. We implemented our method using Python (version 3.9.16) and executed the experiments in a computer with 96 GB memory and Intel Core i9-10900 (2.80 GHz) processor (Our implementation is available from the Github repository [25]). For comparison with the conventional method, we used ARX [26-28], which is recognized as a de facto standard, open-source software for anonymizing personal data.

Our ρ-RDP framework per se releases ρ-safe data records, irrespective of noise types. In addition, it incurs reasonable processing overhead of $O(N \cdot M)$ complexity. Therefore, in this section, we focus on evaluating the data utility of released data via our methods and the conventional data perturbation method with random noise - the basic RDP in Definition 1 (Note that the data records perturbed via the conventional method cannot guarantee our proposed privacy measure "ρ-safeness"). We also have included k-anonymity for comparison with our proposed methods to evaluate data utility between the generalization-based approach and the perturbation-based approach

Firstly, we compare statistics of original and perturbed relations using synthetic data with various settings (N = 1,000 and 10,000; M = 10 and 20; ρ = 0.01 and 0.1). The synthetic data utilized in our first and second evaluations is generated based on a uniform distribution $U(0, 1)$. Table 1 shows the results of statistics obtained from the conventional method, k-anonymity with $k = 3$ and $k = 5$ and our proposed method. The results are averaged over 20 trials, where the measures we use are mean absolute error (MAE), variance of absolute errors (VAE), and Jensen-Shannon distance (JSD) between the original and de-identified data distributions $JSD(r_{i*}, r'_{i*})$. JSD provides a means of quantifying the similarity between two probability distributions. When computing JSD, we bound the perturbed values into range $[0, 1]$. Within Table 1, the minimum values are highlighted in bold, meaning that the method associated with these bolded values exhibits the closest similarity to the original data in terms of statistics.

Table 1. Statistics comparison based on synthetic data

(N, M)	Conv.	Uni.	Lap.	Uni.	Lap.	k-anon	k-anon
Stats.	**method**	$\rho = 0.01$	$\rho = 0.01$	$\rho = 0.1$	$\rho = 0.1$	$k = 3$	$k = 5$
$(1{,}000, 10)$							
MAE	0.5004	0.5002	**0.1004**	0.4997	0.1440	0.1525	0.1817
VAE	0.0834	0.0836	**0.0099**	0.0838	0.0120	0.0152	0.0137
JSD	0.3458	0.3458	**0.1111**	0.3458	0.1483	0.1750	0.1799
$(10{,}000, 20)$							
MAE	0.5001	0.5000	**0.1004**	0.5000	0.1390	0.1910	0.1984
VAE	0.0833	0.0833	**0.0100**	0.0833	0.0115	0.0166	0.0159
JSD	0.3496	0.3495	**0.1140**	0.3493	0.1448	0.1908	0.1952

The minimum values are in bold.

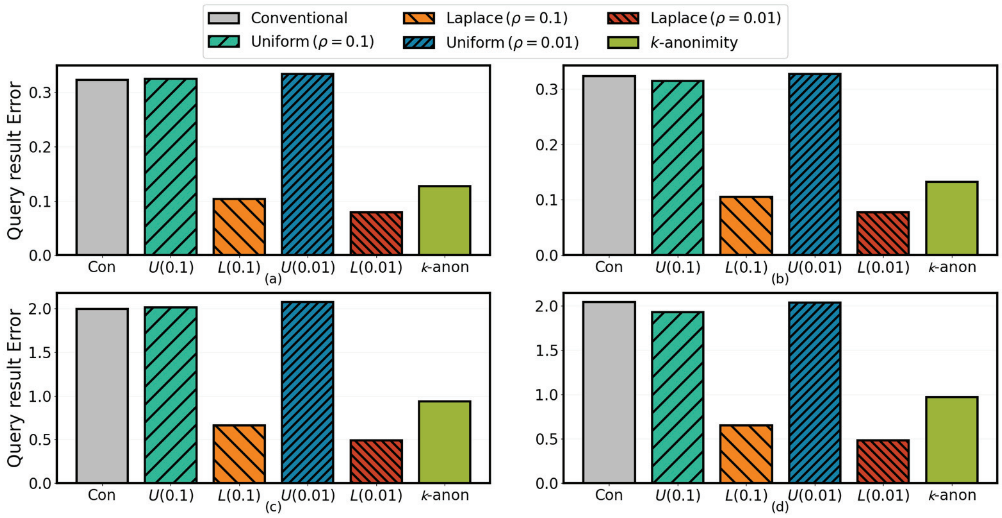


Fig. 4. Linear query result comparison based on synthetic data. (ρ, N, M) = (0.01, 1000, 10), k = 5, (a), (b), (c) and (d) corresponding to L_1– L_4, respectively.

Since shuffling the data records after the perturbation ensures the plausible deniability to individuals [24], we need not care coincidental matches, that is $r_i' = r_j (i \neq j)$. So, the privacy risk measures used in [14,18], such as hitting rate or distance to closest record (DCR), are not considered here.

Secondly, we compare the performance of linear queries. The linear query involves calculating a linear combination of the counts within a data vector, encompassing various aggregation tasks such as counting and summing. Since linear query has the capability to represent a wide range of common aggregation queries, it is one of the most popularly used query forms in data analyses [5]. We use four queries L_1–L_4 as follows, and Fig. 4 depicts the query result error compared with original data.

$(L_1)\ r_{i,1} + r_{i,2} + r_{i,3} \dots + r_{i,M-1} + r_{i,M}$

$(L_2)\ r_{i,1} - r_{i,2} + r_{i,3} \dots + r_{i,M-1} - r_{i,M}$

$(L_3)\ r_{i,1} + 2 \cdot r_{i,2} + 3 \cdot r_{i,3} \dots + (M-1) \cdot r_{i,M-1} + M \cdot r_{i,M}$

$(L_4)\ r_{i,1} - 2 \cdot r_{i,2} + 3 \cdot r_{i,3} \dots + (M-1) \cdot r_{i,M-1} - M \cdot r_{i,M}$

In Table 1 and Fig. 4, the conventional perturbation methods with random noise do not provide accurate data compared with our methods, since they do not add noise adequately for the given privacy requirement. In the case of k-anonymity, it has shown a comparable level of accuracy to our proposed methods. With regards to our methods, the Laplace noise method outperforms the uniform noise method. This is because the amount of noise via the former is adequate to satisfy the given privacy requirement, not too much. Also, the distribution of source data is preserved well especially by the Laplace method. Obviously, perturbation with small ρ provides more accuracy.

Lastly, we evaluate the data utility w.r.t. aggregation query processing using a real dataset (UCI Adult data [29]) in Table 2. The Adult dataset consists of information for 32,561 individuals. Each record in the dataset contains a total of 15 attributes: six numeric attributes and nine categorical attributes. In this experiment, only numeric attributes are perturbed where the domain is set as [*min-value*, *max-value*] for each attribute. And with respect to the case of k-anonymity, certain numeric attributes, including *age*, *education-num*, *capital-gain*, *capital-loss* and *hours-per-week*, are determined as quasi-identifiers and the data was anonymized using the data anonymization tool, ARX [26]. The results of the SQL queries are displayed in Table 2. The bolded values indicate the highest similarity with the original data.

We use the following SQL queries A_1–A_6.

(A_1) select avg(*age*) from adult where *race* = 'White' and *sex* = 'Female';

(A_2) select count (*) from adult where *marital-status* = ‘Divorced’ and *workclass* = 'Private' and *age* > 40;

(A_3) select avg(*hours-per-week*) from adult where *education* = 'Doctorate' and *occupation* = 'Prof-specialty';

(A_4) select count (*) from adult where *education-num* < 10 and *age* > 30 and *relationship* = 'Unmarried';

(A_5) select avg(*capital-loss*) from adult where *education* = 'Bachelors' and *age* < 40;

(A_6) select count (*) from adult where *education-num* > 9 and *hours-per-week* < 45 and *salary* = '>50K'.

Table 2. Aggregation query result comparison based on real data

Q. No.	Orig. data	Conv. method	Uni. $\rho = 0.01$	Lap. $\rho = 0.01$	Uni. $\rho = 0.1$	Lap. $\rho = 0.1$	k-anon $k = 3$
A_1	36.82	36.28	37.02	**36.83**	36.97	36.85	34.77
A_2	1652	1607	1561	1695	1587	**1663**	721
A_3	48.05	45.15	50.46	45.72	44.36	47.36	**48.72**
A_4	1370	941	1010	**1375**	967	1209	561
A_5	101.4	60.9	173.4	130.4	**111.3**	87.0	2178
A_6	2998	2242	2294	**3116**	2270	2789	2497

The highest similarity with the original data is in bold.

In case of the third experiment, the performance results of our and conventional methods are not clearly distinguishable, since the queries return aggregation results and all methods are based on probability distributions with mean '0'; however, our methods (especially with the Laplace noise) outperform the conventional one for the same reason as in the first and second experiments. When applying k-anonymity, it is observed that certain attributes, experience a substantial loss of data, causing some significantly different values from original query result compared to our methods. For example, in query A_5, values in the attribute *capital-loss* undergoes significant suppression, leading to notable differences between the query result from the anonymized data and the original data.

5. Conclusion

In the paper we proposed a syntactic privacy criterion "ρ-safeness" for data perturbation, through which relational data can be perturbed in a quantitatively controlled, privacy-preserving way. To the best of our knowledge, there were no perturbation-based data anonymization methods supporting the quantitative control of privacy. And we devised two perturbation methods that guarantee the given privacy criterion ρ. Through experiments with synthetic and real data, we empirically evaluate the performance of proposed methods in comparison with the conventional generalization-based method. Based on the results, we found that the proposed method effectively anonymizes the relational data in an efficient manner.

For future work, we will consider datasets that contain not only numeric attributes but also categorical attributes. To address this challenge, we need develop practical and reasonable scoring functions to quantify the distance between categorical attribute values. Furthermore, we will study on the privacy-constrained perturbation using multiplicative noises which are known to preserve the locality of data records and hence are useful for data mining tasks such as clustering, classification and regression [4].

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (No. IITP-2023-2020-0-01819, IITP-2021-0-00634), and the National Research Foundation of Korea (No. NRF-2020R1A2C2013286, NRF-2021R1A6A1A13044830).

References

[1] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, San Diego, CA, USA, 2003, pp. 211-222. https://doi.org/10.1145/773153.773174

[2] Y. Zhu and L. Liu, "Optimal randomization for privacy preserving data mining," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, 2004, pp. 761-766. https://doi.org/10.1145/1014052.1014153

[3] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, Tokyo, Japan, 2005, pp. 193-204. https://doi.org/10.1109/ICDE.2005.8

[4] C. C. Aggarwal and P. S. Yu, *Privacy-Preserving Data Mining: Models and Algorithms*. New York, NY: Springer, 2008. https://doi.org/10.1007/978-0-387-70992-5

[5] C. Li, "Optimizing linear queries under differential privacy," Ph.D. dissertation, University of Massachusetts Amherst, Amherst, MA, USA, 2013.

[6] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557-570, 2002. https://doi.org/10.1142/S0218488502001648

[7] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, article no. 3-es, 2007. https://doi.org/10.1145/1217299.1217302

[8] V. T. Gowda, R. Bagai, G. Spilinek, and S. Vitalapura, "Efficient near-optimal t-closeness with low information loss," in *Proceedings of 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Cracow, Poland, 2021, pp. 494-498. https://doi.org/10.1109/IDAACS53288.2021.9661004

[9] C. Dwork, "Differential privacy," in *Automata, Languages, And Programming*. Heidelberg, Germany: Springer, 2006, pp. 1-12. https://doi.org/10.1007/11787006_1

[10] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 84, no. 1, pp. 3-37, 2022. https://doi.org/10.1111/rssb.12454

[11] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, "Differentially private data publishing and analysis: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1619-1638, 2017. https://doi.org/10.1109/TKDE.2017.2697856

[12] H. Jiang, J. Pei, D. Yu, J. Yu, B. Gong, and X. Cheng, "Applications of differential privacy in social network analysis: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 108-127, 2023. https://doi.org/10.1109/TKDE.2021.3073062

[13] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao,"). PrivBayes: private data release via Bayesian networks," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, pp. 1-41, 2017. https://doi.org/10.1145/3134428

[14] P. H. Lu, P. C. Wang, and C. M. Yu, "Empirical evaluation on synthetic data generation with generative adversarial network," in *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, Seoul, Republic of Korea, 2019, pp. 1-6. https://doi.org/10.1145/3326467.3326474

[15] J. Fan, T. Liu, G. Li, J. Chen, Y. Shen, and X. Du, "Relational data synthesis using generative adversarial networks: a design space exploration," 2020 [Online]. Available: https://arxiv.org/abs/2008.12763.

[16] Financial Services Commission, "Guidelines for Financial Data Pseudonymization and Anonymization," 2022 [Online]. Available: https://www.fsec.or.kr/bbs/detail?menuNo=246&bbsNo=6484.

[17] Korean Law Information Center, "Personal Information Protection Act," 2023 [Online]. Available:

https://www.law.go.kr/LSW/lsInfoP.do?chrClsCd=010203&lsiSeq=142563&viewCls=engLsInfoR&urlMode=engLsInfoR/1000#0000.

[18] PWS Cup 2018 [Online]. Available: https://www.iwsec.org/pws/2018/cup18.html.

[19] M. Rahman, M. K. Paul, and A. S. Sattar, "Efficient perturbation techniques for preserving privacy of multivariate sensitive data," *Array*, vol. 20, article no. 100324, 2023. https://doi.org/10.1016/j.array.2023.100324

[20] *Privacy enhancing data de-identification terminology and classification of techniques*, ISO/IEC 20889:2018, 2018.

[21] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proceedings of the 31st VLDB Conference,* Trondheim, Norway, 2005, pp. 901-909. https://dl.acm.org/doi/10.5555/1083592.1083696

[22] D. Wang, B. Guo, and Y. Shen, "Method for measuring the privacy level of pre-published dataset," *IET Information Security*, vol. 12, no. 5, pp. 425-430, 2018. https://doi.org/10.1049/iet-ifs.2017.0341

[23] C. K. Liew, U. J. Choi, and C. J. Liew, "A data distortion by probability distribution," *ACM Transactions on Database Systems (TODS)*, vol. 10, no. 3, pp. 395-411, 1985. https://doi.org/10.1145/3979.4017

[24] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, TX, USA, 2000, pp. 439-450. https://doi.org/10.1145/342009.335438

[25] Github, "rho_RDP," 2023 [Online]. Available: https://github.com/jXXXXDy/rho_RDP/tree/main.

[26] F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn, "Flexible data anonymization using ARX: current status and challenges ahead," *Software: Practice and Experience*, vol. 50, no. 7, pp. 1277-1304, 2020. https://doi.org/10.1002/spe.2812

[27] C. E. Jakob, F. Kohlmayer, T. Meurers, J. J. Vehreschild, and F. Prasser, "Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19," *Scientific Data*, vol. 7, article no. 435, 2020. https://doi.org/10.1038/s41597-020-00773-y

[28] A. C. Haber, U. Sax, F. Prasser, and NFDI4Health Consortium, "Open tools for quantitative anonymization of tabular phenotype data: literature review," *Briefings in Bioinformatics*, vol. 23, no. 6, article no. bbac440, 2022. https://doi.org/10.1093/bib/bbac440

[29] UCI Machine Learning Repository, "Adults dataset," 1996 [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Adult.

Deokyeon Jang **https://orcid.org/0009-0006-8912-0421**

He received the B.S. degree from the Department of Computer Science and Engineering, Korea University, Seoul, South Korea, in 2023, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering. His research interests include data privacy and machine learning for databases.

Minsoo Kim **https://orcid.org/0000-0003-3450-9721**

He received the B.S. degree in computer and information science from Korea University, Sejong, South Korea, in 2015, and the Ph.D. degree with the Department of Computer Science and Engineering from Korea University, Seoul, South Korea, in 2023. His research interests include array database and distributed/parallel processing of large-scale data.

Yon Dohn Chung **https://orcid.org/0000-0003-2070-5123**

He received the B.S. degree in computer science from Korea University, Seoul, South Korea, in 1994, and the M.S. and Ph.D. degrees in computer science from KAIST, Daejeon, South Korea, in 1996 and 2000, respectively. He was an Assistant Professor with the Department of Computer Engineering, Dongguk University, Seoul, from 2003 to 2006. He joined as a faculty member of the Department of Computer Science and Engineering, Korea University, in 2006, where he is currently a professor. His research interests include database systems, spatial databases, and data privacy.