Differential Privacy and k-Anonymity-based Privacy Preserving Data Publishing Scheme with Minimal Loss of Statistical Information

Abdul Majeed *Member, IEEE*, and Seong Oun Hwang *Senior Member, IEEE*

***Abstract*—Though anonymization mechanisms have made huge progress in fostering the secondary use of data, it is still very challenging to obtain adequate knowledge from anonymized data while preserving privacy. Most existing mechanisms anonymize entire sections of data, and fail to maximally preserve the structure/values of real data. Consequently, the performance of those mechanisms and the output (i.e., the anonymized data) remain problematic in real-life scenarios due to the extensive and unneeded anonymization applied. To address these issues, we propose and implement a hybrid (differential privacy and k-anonymity) anonymization scheme that produces supreme-quality anonymized data that offers knowledge similar to real data without compromising privacy. Specifically, we implement a pair of algorithms that divide the dataset into privacy-violating and non-privacy-violating partitions. Afterward, in a non-privacy-violating partition, a relaxed privacy budget ϵ is applied to numerical attributes, but most of the categorical attributes are retained (as is) for informative analysis. In privacy-violating partitions, fewer changes are applied to the data by using a reasonable value for ϵ and by exploiting the diversity in sensitive information. Experiments are conducted on three real-life datasets to prove the feasibility of our scheme for futuristic AI applications. Compared with state-of-the-art (SOTA) methods, our scheme preserves 60.81% of the originality in the anonymized data. The privacy risks are reduced by 20.05%, and utility is enhanced by 54.01% and 15.33% based on information loss and accuracy metrics. Furthermore, the time overhead is 3.13$\times$ lower than the SOTA methods.**

***Index Terms*—privacy preserving data publishing, differential privacy, k-anonymity, statistical information, privacy, utility**

I. Introduction

RECENTLY, personal data has replaced oil as the most economically desirable resource in the world, and therefore, most companies are striving to make the most of the economic rewards from it by using advanced data mining tools. Data are no longer just raw materials, but products with tremendous opportunities for profit [1]. Data is becoming a powerful source in augmenting the performance of many real-world, data-driven services such as healthcare, epidemic mitigation, and decision-making. Although data has become an economic resource, there is a growing demand for fair, and responsible use of data in the AI era [1]. Besides, privacy preservation is the main barrier to exploiting the full potential of personal data [2]. Due to privacy concerns, most organizations are not willing to share their data, and therefore, extracting valuable knowledge is the stuff of dreams. The COVID-19 pandemic has also shown that privacy is a major bottleneck when it comes to handling personal data [3]. The recent European law (i.e., GDPR) put special emphasis on the responsible use of data without sacrificing privacy [4]. Under such law, all firms that deal with any kind of personal data are required to use privacy protection technology to stay legally compliant.

Well-known methods for preserving privacy in personal data are syntactic and semantic [5]. The former, such as k-anonymity [6], ℓ-diversity [7], t-closeness [8], and their latest versions [9], preserve privacy by generalizing the data. The latter, such as differential privacy (DP) [10] and its improved versions [11], preserve privacy by adding noise to the data. Many improved versions of these methods have demonstrated effectiveness in anonymizing big data with a greater balance of both utility and privacy [12]–[14]. Although both methods (syntactic and semantic) help in data publishing, there are five major problems to these methods from the perspective of responsible data science (RDS)[1].

- Most syntactic methods anonymize all parts of data, which makes knowledge extraction harder, and the exceedingly anonymized data cannot be used in AI applications. In addition, poorly anonymized data increases the difficulty for data consumers/analysts.
- Most syntactic methods lose truthfulness in data during anonymization by either using suppression or wide generalization intervals, leading to concept-/data-drift issues in ML applications.
- Most semantic methods apply a fixed ϵ value to the entire dataset, which adds more noise, leading to lower usability in data-driven applications [15].
- Most semantic methods add excessive noise to values in a minor population, which can hinder knowledge discovery from all perspectives of the data.
- Both methods do not identify and abstract the privacy-violating and non-privacy-violating patterns imperative to preserving privacy in data sharing [16].

The major contributions of this paper are given as follows.

- We explore invisible issues with syntactic and semantic methods when it comes to the quality of data in the context of RDS[1] and data-hungry AI applications. By implementing a new anonymity scheme, we identify opportunities to amalgamate DP and k-anonymity to safeguard the privacy and to produce supreme-quality

This paper was produced by the Information Security and Machine Learning Research Group. They are in Seongnam, South Korea.

Manuscript received April 19, 2021; revised August 16, 2021.

[1]https://redasci.org/

data that enable informed decision-making. Our scheme successfully resolves the above-cited challenges in the existing methods.

- Our scheme divides data into two partitions: non-privacy-violating and privacy-violating. It retains most values in their original form in the former partition, and performs the minimal necessary anonymization in the latter by applying a k-member clustering that incorporates similarity and diversity in attributes, and that produces compact and diverse clusters while satisfying k-anonymity criteria.
- We implement an ML-based pattern-computing method that can assist in identifying attributes that encompass the possibility of non-privacy-violating and privacy-violating patterns, whereas prior methods do not identify such patterns, and lead to extensive anonymization and poor privacy guarantees.
- We developed a hybrid data transformation scheme, which applies a relaxed ϵ to numerical attributes and lower-level generalization to categorical attributes to create a very close representation of the original data.
- We performed exhaustive experiments on three real-life datasets to prove the technical effectiveness of our scheme for futuristic AI applications. When anonymized data is intended to be used in AI applications, it is necessary to maximally preserve the statistical information (e.g., all feature values and their distributions) in it to prevent harmful/biased decisions. To this end, our scheme is handy as it can preserve higher truthfulness in the anonymized data and better retain the semantics of real data, leading to correct data use in AI applications.
- The main novelty of our scheme lies in the preprocessing of data through the identification of pattern-friendly QIDs via ML, classification of data into privacy-violating and non-privacy-violating parts that allows distinct values of ϵ rather than fixed ones, anonymization of some parts only rather than entire dataset, preserving most parts of data in their original form for higher utility along with the strong privacy guarantees, and least computing overheads.

The rest of this paper is organized as follows. Section II provides the system model and formulates the problem. Section III introduces our hybrid scheme in detail. Section IV demonstrates the results and comparison on three different datasets. Section V concludes this paper.

II. System Model and Problem Formulation

In this work, we consider a generic data publishing scenario in which five actors are involved (record owners, database owners, database publishers, analysts, and adversaries) as shown in Figure 1. Record owners provide their data in the form of tuples/records to database owners. The data collected from N record owners are orchestrated by database owners. Subsequently, the data in the anonymized form is either directly released by the database owners or by database publishers and is used by analysts for knowledge discovery. The knowledge derived from the released data is used to improve real-world applications (e.g., healthcare). In this system, there exists an adversary whose goal is to jeopardize record owner privacy in the published data by linking auxiliary information. Each of the N record owners in the system can have distinct values for his/her attributes. For example, if $N = 20$, 18 record owners might have White as their race attribute, and two might have Black as that attribute. In this situation, we can divide race information into two parts, privacy-violating (*Black*) and non-privacy-violating (*White*). Our goal is to devise a privacy-preserving scheme for data sharing that protects privacy-violating information from adversaries while effectively releasing non-privacy-violating data (as is) for analytics. We assume that the dataset (D) to be anonymized has been gathered from pertinent record owners, and rows in D associate to the real-life persons with their demographics, both quasi-identifiers (QIDs) and sensitive attributes (SAs). Any real-life D containing QIDs and SAs can be anonymized with our algorithm.

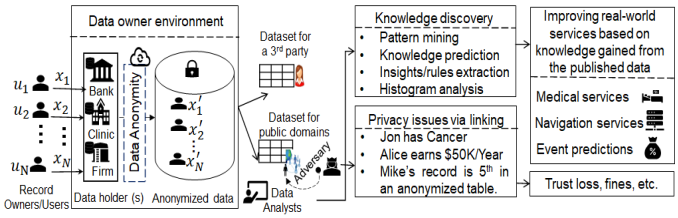


Fig. 1. Schemetic of system model considered in this paper.

Threat model: This work assumes that most entities in the system are honest. They perform only the required actions, and help accomplish the desired tasks from data releases. However, some data analysts can behave like an adversary and jeopardize privacy. Although we remove all kinds of directly identifiable information from the data, QIDs can still be acquired from various sources and can be linked to identify people [17]. Hence, our algorithm is vulnerable to identity and corresponding SA disclosure in two ways.

- Adversaries may already have access to some QIDs and might attempt to figure out the remaining QIDs. For example, an adversary who knows the sex and age of a person might try to identify his/her zip code.
- Adversaries may know the entire tuple (e.g., all the QIDs) of a record owner in advance and might also know with a high probability that that person's information is among the released data. Based on this reliable information, he/she essay to obtain the SAs of a target user. For instance, data can include record owners' SAs on monthly income or diseases contracted. If the adversary can somehow identify or associate QIDs correctly, he/she can also determine an SA akin to that record owner.

To this end, we intend to protect personal privacy from these present-day privacy perils that can emerge in privacy-preserving data publishing (PPDP).

Privacy model: Our model uses a generalization and a Laplace mechanism to anonymize data. An overview of the privacy model, including both techniques, is in Figure 2. The definitions for both models are as follows.

Definition 1 (k-anonymity): A sanitized/anonymized dataset, $D^{'}$, obtained from real dataset D adheres to k-anonymous if each record has at least k identical records in each QID group.

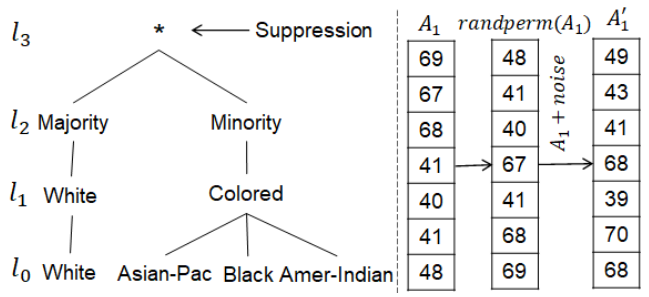


Fig. 2. Privacy model: (left) domain generalization hierarchy (DGH) for the race, and (right) noise addition for age via the laplace mechanism.

Definition 2 (Laplace mechanism): The Laplace mechanism is used to ensure ϵ-DP, where the output of function F on D is in the form of a real-number vector [18]. It injects noise n into every value in the output of $F(D)$ to guarantee ϵ-DP. The n is taken from a Laplace distribution having mean 0 and scale ($sensitivity/scale$).

Design goals and problem formulation: This paper aims to achieve the following three major goals in PPDP.

Originality: Ensures that most QID values in non-privacy-violating partitions are not generalized in the anonymity process and remain as close as possible to D.

Utility: Ensures that $D^{'}$ (anonymized data) retains maximal knowledge to improve real-life services. Concisely, it lowers information loss (IL) and enhances accuracy.

Privacy: Ensures that an adversary having myriad auxiliary information/data cannot match/link QIDs and infer SAs with a higher probability. It guarantees that when an adversary tries to associate a person in $D^{'}$ to any auxiliary data, there will be a link from any record to numerous SA values. Concisely, it prevents linking the SAs of any record.

The main problem to be addressed with our hybrid scheme is formally expressed in Problem 1.

Problem 1. *Given real-world dataset D encompassing various attributes (name, age, sex, race, income/disease, privacy parameter k, privacy budget ϵ), how do we produce an anonymized data $D^{'}$ where (a) $D \subseteq D^{'}$, (b) $D^{'}$ is k-anonymous, (c) $D^{'} \sim D$ (e.g., most parts in $D^{'}$ are highly the same as original D), and (d) $D^{'}$ has exceptional quality (a.k.a. utility) in terms of analytics (i.e., has both significantly low information loss and high accuracy).*

III. Hybrid Anonymity Scheme

In this section, we present the proposed hybrid anonymity scheme in detail. Figure 3 shows the workflow of our scheme. There are four main components: pre-processing, identification of pattern-friendly QIDs in which patterns exist, dividing data into privacy-violating and non-privacy-violating partitions, and applying data transformation (generalization and noise addition) to both partitions. Descriptions of each component are as follows.

Pre-processing of D: Generally, any real-world D can contain four types of attributes concerning record owners: QIDs, SAs, explicit identifiers (EIs), and non-sensitive attributes (NSAs). The definitions and examples of each attribute type are given in Figure 4.

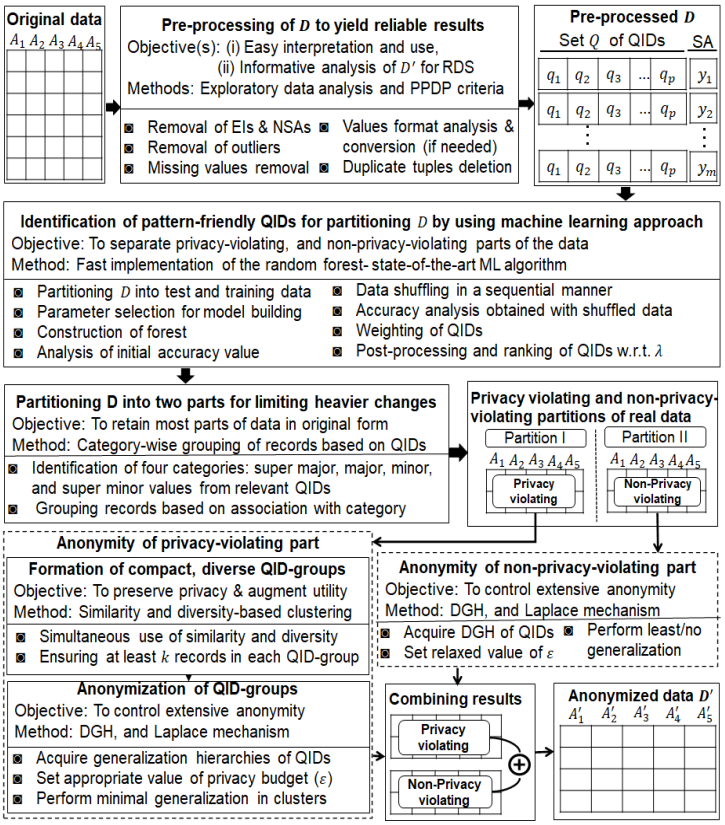


Fig. 3. Workflow of the proposed hybrid anonymity scheme.

After getting D, we apply pre-processing to clean D for further processing via the five steps shown in Figure 3. First, we remove two types of attributes (NSAs and EIs) per the standard routine for PPDP. EIs are removed to lessen identity leakage from $D^{'}$, whereas NSAs have a minimum effect on utility/privacy. It is worth noting that NSAs can have some impact on utility in some cases. For example, the weight has the

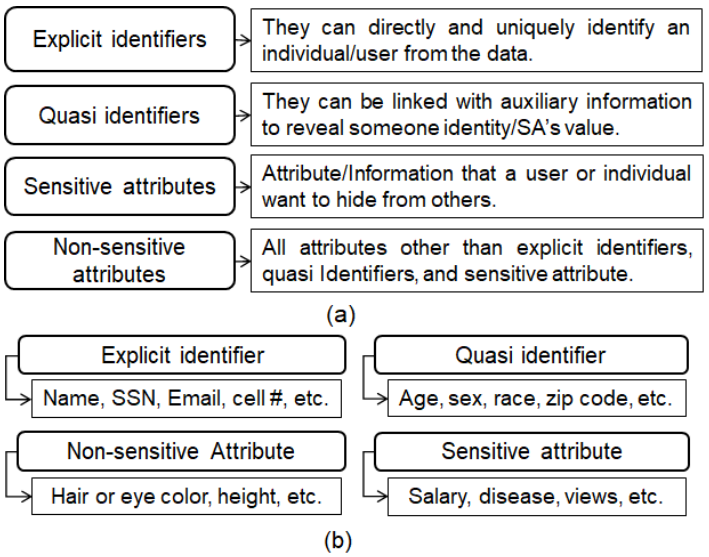


Fig. 4. Attribute types present in the data: (a) definitions, and (b) examples.

least utility when analyzing w.r.t. income or political/religious

views. However, weight can be an important attribute when doctors want to analyze the relationship between illness and weight. Therefore, ample attention is required while removing the NSAs from the data. In our recent work [19], we examined the impact of NSAs on data utility and found that NSAs can yield only very small improvements in accuracy. Therefore, the decision concerning the retention of NSAs in the final anonymized data can be made depending on either the objectives of the data release or the nature of information consumers. In practice, NSAs are usually not collected from individuals at the data collection time. If NSAs are collected then two approaches are traditionally followed: (i) removed from data, or (ii) published as is in the anonymized data. After NSA and EI removal, QIDs and SAs are restructured (an SA is placed at the last column usually). The final structure of D becomes $D\{Q, Y\}$, where $Y = \{y_1, y_2, \ldots, y_m\}$ and $Q = \{q_1, q_2, \ldots, q_p\}$. An example of D's common template is given in (1):

$$D_{U,A} = \begin{pmatrix} u_i & q_1 = age & q_2 = sex & q_p = race & Y = disease \\ 1 & 39 & F\cdots & Black & Cancer \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 9000 & 37 & M\cdots & White & HIV \end{pmatrix} \quad (1)$$

where each row and column provide complete and partial (one item) information about a person, respectively.

Because D is curated from diverse sources and people, it can contain outliers (i.e., undesirable values in some columns; for instance, age value $\neq$ 600, but is highly likely 60). Outliers can corrupt analytical results and therefore should be eliminated. To remove outliers, we perform $min-max$ analysis and create visual plots to ensure whether each attribute value is within a valid range or not. Afterward, we remove records with incomplete information. Furthermore, we check the consistency of attributes' values w.r.t data type (numerical, categorical). Redundant records are discarded at this stage to lessen computing complexity. In the end, we transform some QID types (discrete $\rightarrow$ numerical) by using *key-value* concept. For example, sex can be transformed as either discrete or numerical via two key-value pairs like $[0, F]$ and $[1, M]$. Furthermore, in some cases, data are not in the desired format owing to the direct scanning of documents. Hence, the values' format can be transformed accordingly to perform the relevant operations. By adopting the above-cited steps, a good quality D is curated for additional processing.

Identifying pattern-friendly QIDs from D: Identifying QIDs that can possess privacy-violating and non-privacy-violating patterns can contribute to preserving both privacy and utility [16]. To identify such QIDs, we employ a machine learning technique named random forest (RF) [20]. RF is a highly reliable ML method that has shown remarkable achievements in accomplishing prediction/classification tasks in many domains. In this work, we build an RF model with a data-shuffling strategy to figure out the desired QIDs. Specifically, we build the model with D using QIDs as predictors and SAs as target classes and obtain a reference accuracy value. Subsequently, we shuffle the data (one QID at a time) in each iteration and build the RF model again. Owing to the shuffled data, the new accuracy value can be either higher or lower than the reference accuracy. If a QID has many similar values, the accuracy does not change much, and such a QID can have a higher possibility of containing privacy-violating and non-privacy-violating patterns. With the help of the above method and with some minor post-processing, we can correctly identify QIDs that possibly have the patterns. The procedure for identifying pattern-friendly QIDs is given in Figure 5.

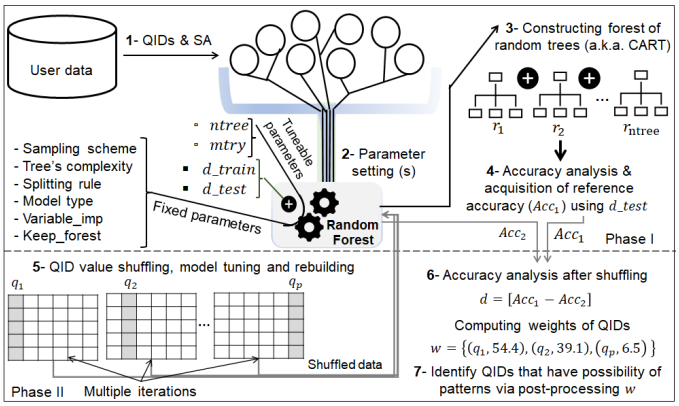


Fig. 5. The proposed method for identifying pattern-friendly QIDs in D.

Dividing D into privacy and non-privacy violating partitions: We partition D into two parts based on the pattern-friendly QIDs identified in the previous step. We identify four types of values (super major, major, minor, super minor) from relevant QIDs, and then partition D. We explain the mathematical foundation for dividing D via Example 1.

Example 1. For any pattern-friendly QID, $q_i \in Q$, the unique values $\in q_i$ can lie into one of the four categories (c_1, c_2, c_3, c_4) with very high probability (P), and the difference between the c_1 and c_4 categories, $|c_1 - c_4|$, is large. For a chosen q_i, we determine the unique values and then denote them with $v_1^*, v_2^*, \ldots, v_e^*$. Thus, for each v_l^* (unique value), there prevail many records $f(q_i, v_l^*)$. After finding $f(q_i, v_l^*)$, v_l^* can be mapped to one of the four candidates(e.g., c_1, c_2, c_3, c_4) depending on f where $f = |(v_l^*, q_i)|$. In simple words, if f of v_l^* is very high, it can be mapped to c_1 directly, and if the f is lowest, it is included in c_4. The other two categories are settled based on values of f. Since in pattern-friendly QIDs, the values' distribution is highly imbalanced, each value can be correctly mapped to the respective category with high P. Due to the imbalanced distributions, the discrepancy between c_4 and c_1 is large in most cases. A formal aspect by taking the example of race QID in Q focusing on Example 1 is expressed below:

$$q_i(race) = \begin{bmatrix} v_1^* = white,\ f(white) = 27,816, v_1^* \in c_1 \\ v_2^* = black,\ f(black) = 3,124, v_2^* \in c_2 \\ v_3^* = API,\ f(API) = 1,039, v_3^* \in c_3 \\ v_4^* = AIE,\ f(AIE) = 311, v_4^* \in c_4 \\ v_5^* = other,\ f(other) = 271, v_5^* \in c_4 \end{bmatrix}$$

In the above formalization, race value AIE $\in c_4$ rather than c_3 due to relatively less representation (e.g., # of records) of it in the data. For example, if a k value close to 300 is used to create classes from data having AIE in each record, there

will be hardly one class that can be generated from c_4, which indicates that it is a super minor value. In contrast, a similar k value will result in more than 3 classes from c_3 because it is a minor value. The major difference between c_3 and c_2 is the # of records and their dilution chances when anonymity is applied to them. Since c_2 has a relatively higher record than c_3, and therefore, the chances of its dilution are comparatively less than the c_3, and vice versa. It is worth noting that in some real-world datasets, the cardinality of QIDs can be low, and therefore, only a few categories among four can be enforced depending upon the scenario.

In application, the QID can be categorized based on their availability at external sites, privacy and utility requirements, and the association with an individual's identity [21]. In our scheme, we have chosen a subset of attributes from data as QIDs just like previous approaches do. Later, pattern-friendly QIDs were categorized through the implementation of the RF method discussed in the former step. From experiments, we found that a particular QID can be categorized as pattern friendly, when the respective QIDs has one/two value that makes up more than 80/85% of the data, and the rest of the data constitute multiple least frequency values. These findings enabled us to apply anonymity to some parts of the data only, leading to significantly better results than the SOTA methods.

A similar process is applied to all pattern-friendly QIDs, and D is partitioned based on major and minor values. The procedure applied to partition D is given in Figure 6.

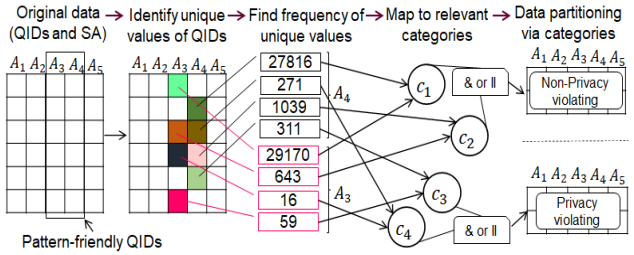


Fig. 6. The proposed method for partitioning D into the two categories.

The main innovation of our scheme lies in the pre-processing of data through the identification of pattern-friendly QIDs. However, even if a QID is considered pattern-friendly, there is still a certain risk of privacy disclosure associated with it. Through experiments, we found that in the adult dataset, the country QID is highly pattern-friendly, but there is still an 8.84% risk of privacy disclosure associated with it. However, the risk of privacy disclosure reduces in subsequent steps by applying the generalization operation.

Anonymization of data in both partitions: In the last step, D is anonymized to yield $D^{'}$. We add noise (with a relaxed ϵ) to numerical QIDs, and apply the least generalization to discrete QIDs. Algorithm 1 details the process of numerical QID anonymization located at index 1 of P_1. In Algorithm 1, the relevant partition, ϵ, and the sensitivity are provided as input, and partially anonymized data is returned as output. The keys to anonymity are steps 6 and 7. Further explanation of algorithm 1 that was applied to numerical QIDs is given in Figure 7. Referring to Figure 7, the QIDs are first split into numerical (N) and categorical (C), respectively. Afterward, noise is generated with optimal value of ϵ, and added to numerical QIDs only. In noise curation, it is vital to choose a suitable ϵ to preserve privacy and utility. In our algorithm, we have chosen optimal values of ϵ for each partition.

Algorithm 1 $CreateAnonymizedDataNumerical(P_1, \epsilon, \triangle f)$

Require: $P_1, \triangle f, \epsilon$
Ensure: $D^{''}$
1: $D^{''} \leftarrow \emptyset$ ▷ $D^{'}$ will store partially anonymized data.
2: $[X, n] \leftarrow size(P_1)$
3: $dsrand \leftarrow P_1(randperm(X), :)$ ▷ Shuffling of data.
4: $scale \leftarrow {}^{\triangle f}/_{\epsilon}$
5: **for** $j = 1\,to\,|P_1|$ **do**
6: $laplace = round(randlap(0, scale))$ ▷ Get noise
7: $dsrand(j, 1) = dsrand(j, 1) + laplace$ ▷ Add noise
8: **end for**
9: $t = dsrand(:, 5)$, $x = dsrand(:, 1:4)$
10: $D^{''} \leftarrow [D^{''} \cup x\ t]$
11: return $D^{''}$

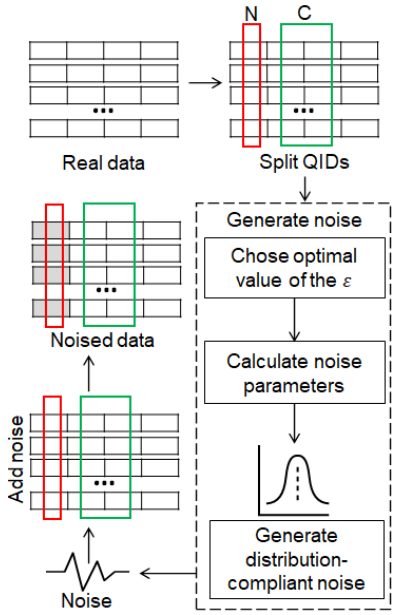


Fig. 7. The explanation of algorithm applied to numerical QIDs.

The ϵ given to each partition was 3.0 (privacy-violating) and 5.0 (non-privacy-violating), respectively. Categorical QIDs can be anonymized using Algorithm 2 with the help of generalization hierarchies. In Algorithm 2, first, the compact clusters are formed using k-anonymity criteria with the help of k-member clustering; diversity div is computed from an SA in each cluster and is compared with the threshold, and then, anonymity is performed considering the div values.

Further explanation of algorithm 2 that was applied to categorical QIDs is given in Figure 8. Referring to Figure 8, the QIDs are first split into numerical (N) and categorical (C), respectively. Afterward, generalization taxonomies/hierarchies for categorical QIDs are either acquired (if pre-built) or created by analyzing the domain values of the respective QIDs. For

Algorithm 2 $CreateAnonymizedDataCategorical(P_2, T, k)$

Require: P_2, T, k
Ensure: $D^{'}$
$D^{'} \leftarrow \emptyset$ ▷ $D^{'}$ will store the anonymized data.
$ClusNum \leftarrow \lceil |P_2|/k \rceil$ ▷ Find total # of classes.
$R \leftarrow k - member(ClusNum, k, P_2)$ ▷ Create clusters.
for $i = 1\,to\,|R|$ **do**
  Compute diversity div of SA values in R_i
  $div(R_i) \leftarrow -\sum_{i=1}^{|Y|}[(p_i) \times ln(p_i)$
  if $(div(R_i) \geq T_{div})$ **then** ▷ Case-I
    for $j = 1\,to\,|Q|$ **do**
      Acquire T_{q_j} ▷ Least generalization case
      $q_j^* \leftarrow Anonymize(q_j, T_{q_j})$, delete q_j from R_i
      $R_i^{**} \leftarrow R_i^* \cup q_j^*$ ▷ Temp. var. to hold results.
      Repeat same process for q_{j+1} to $|Q|$
      return R_i^{**}
    end for
  else if $(div(R_i) < T_{div})$ **then** ▷ Case-II
    for $j = 1\,to\,|Q|$ **do**
      Acquire T_{q_j} ▷ Average generalization case
      $q_j^* \leftarrow Anonymize(q_j, T_{q_j})$, delete q_j from R_i
      $R_i^* \leftarrow R_i^* \cup q_j^*$ ▷ Temp. var. to hold results.
      Repeat same process for q_{j+1} to $|Q|$
      return R_i^*
    end for
  end if
  $D^{'} \leftarrow D^{'} \cup \{R_i^*, R_i^{**}\}$
end for
return $D^{'}$

example, sex can have two values (e.g., males and females). Afterward, the number of levels for generalization taxonomy is determined, and a whole taxonomy in the form of a tree is created. The leaf nodes are real values of QIDs, and the other levels are generalized values. It is important to note that lower levels are suitable for better utility, and vice versa. In the anonymization of QIDs, it is vital to choose a suitable level from the taxonomy to effectively balance both privacy and utility. In our algorithm, we have chosen suitable levels for anonymization in each partition by exploiting the diversity of SA. Algorithms 1 and 2 both assist in anonymizing D in both partitions with the fewest changes.

IV. Experimental Evaluation

In this section, we discuss the results attained from exhaustive experiments on three real-life benchmark datasets.

A. Description of datasets used in experimentation

We performed extensive tests on three datasets: Adults (**A**) [22], Bkseq (**B**) [23], and Careplans (**C**) [24]. These datasets contain a variety of private and public information. We utilized multiple QIDs and SAs in the experiments and removed other non-QID attributes from each D. A concise overview of all three D is in Table I. Pre-processing was rigorously applied to all datasets before experiments. Hereafter, we use the concise form to refer to each dataset (i.e., A, B, and C).

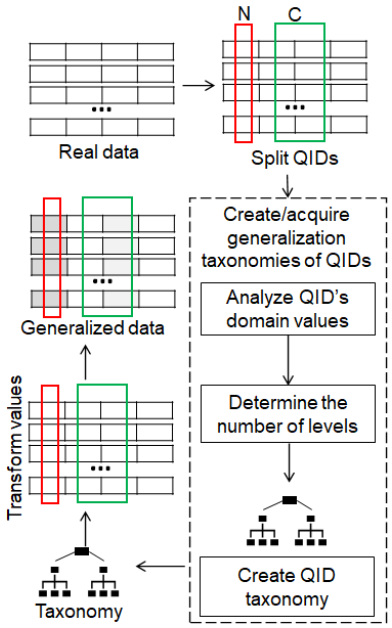


Fig. 8. The explanation of algorithm applied to categorical QIDs.

TABLE I
DETAILS OF THE DATASETS USED IN EXPERIMENT EVALUATIONS.

D	$\|D\|$	QID (Distinct values)	SA Name
A [22]	32,561	Country (41), Age (74), Sex (2), Race (5)	Salary/Income
B [23]	16,160	Age (30), Sex (2), Weight (30)	Medical test results
C [24]	12,352	Race (5), Relationship (3), Sex (2), State (14), Zip code (291)	Healthcare_coverage

B. Implementation setup and comparison criteria

We implemented the scheme on a computer with Intel Core i5-3320M CPU with @ 2.60GHz clock speed running Windows 10 with 8GB RAM. Implementation was done using two software packages: Matlab version (R2022a) and RTools, R version 4.0.0 with the help of built-in packages. The other libraries employed in finding pattern-friendly QIDs were RF [2] and ranger[3] (a fast RF implementation). Descriptions of salient parameters/variables employed in RF are in Table II. Default values of some parameters (sampling scheme, sample size, etc.) were used. After experimentation with RF and post-processing, the following pattern-friendly QIDs were determined from each D: (A: *race, country*), (B: *gender*), and (C: *race, state*). We re-analyzed these findings by checking each QID's real values' domain and distribution in D. The validation results verified reliability.

Baselines and evaluation metrics: We compared the results from our scheme with three SOTA algorithms, k-anonymity [6], SVD3DR [9], and RKA [11]. To the best of the authors knowledge, these are the only recent and relevant SOTA algorithms in this line of work. For fair comparison and assessment, we chose baselines that either strictly follow k-anonymity or do not strictly follow k-anonymity. We also compared our scheme with the RKA method [11], which is also a hybrid method (e.g., uses k-anonymity along with

[2]https://cran.r-project.org/web/packages/randomForest/
[3]https://cran.r-project.org/web/packages/ranger/

TABLE II
IMPORTANT PARAMETERS/VARIABLES USED TO IDENTIFY PATTERN-FRIENDLY QIDS.

Parameter/variable name	Parameter values (A [22], B [23], C [24])	
	Numerical (A,B,C)	Non-numerical
$\|d_{train}\|$	21,707; 10,774; 8,235	-
$\|d_{test}\|$	10,854; 5,386; 4,117	-
Value of $ntree$	494; 269; 225	-
RF model type	-	Clas. (A, B) & Reg. (C)
Splitting rule, Min node size	-	Impurity,1
Value of $mtry$	4; 3; 5	-

the DP) for data anonymization. Lastly, we performed experiments and comparisons at the dataset level (e.g., using entire anonymized datasets produced with different methods), and therefore, the analysis/comparisons are sound and valid. We generated different versions of $D^{'}$ from all datasets, and compared the performance of these algorithms by using the four metrics– originality of values in $D^{'}$, SA disclosure risk (DR), accuracy (Acc), and information loss (IL) via distortion measure (DM)–as follows:

$$DR = \max_{u_i \in D, y \in Y} Pr(u_i[D] = y | D^{'} \wedge \nu) \quad (2)$$

where ν is background knowledge, and $D^{'}$ denotes anonymous data tuples that correspond to ν. Because the tuple that can likely be exploited by an adversary during an attack is unknown, we considered the worst-case scenario in which any tuple can be exploited to infer an SA of an individual.

$$Acc = {}^{TP+TN}/_{TP+FP+TN+FN} \quad (3)$$

where FN, TP, TN, FP, refer to a false negative, true positive, true negative, and false positive, respectively.

$$DM = \sum_{i=1}^{N} \sum_{j=1}^{p} \frac{l}{l'} \times w_{q_j} \quad (4)$$

where l shows the actual level of generalization, and $l^{'}$ shows the total # of levels in T.

C. Identification of pattern-friendly QIDs from D

In this subsection, we present the results of the random forest (RF)-based implementation and the corresponding pattern-friendly QIDs that were identified in each real-life dataset in Figure 9. The circled QIDs in Figure 9 are pattern-friendly QIDs because of their super major and minor categorizations based on the values. For example, in the Adults dataset, 89.58% of the records had the United States as the person's native country, which is a super major value. Similarly, 85.42% of the tuples contain White as the race value. On the other hand, the Other race value had 271 occurrences (0.0083%) in the Adults dataset, which is super minor. These statistics highlight the greater possibility of finding major patterns in some QIDs (e.g., race and country) that can be identified and released in their original form because the risk of privacy violation is low [16]. By identifying pattern-friendly QIDs, most values can be retained in their original form, which can lessen the burden on data analysts. Furthermore, ample attention can be paid to privacy-violating data (e.g., super minor values) to effectively protect privacy against adversaries.

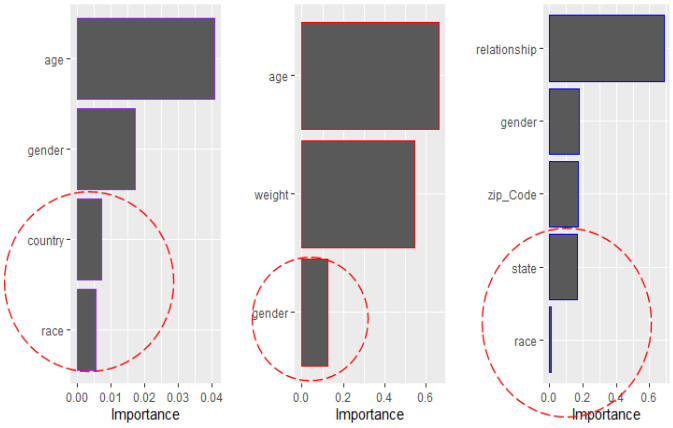


Fig. 9. Pattern-friendly QIDs identified from real-world datasets.

D. Orignality preservation in anonymized data

Anonymization has been successfully applied in many commercial sectors, especially in healthcare, to preserve data privacy while making data broadly available to researchers and data miners. Unfortunately, the data produced by most of the

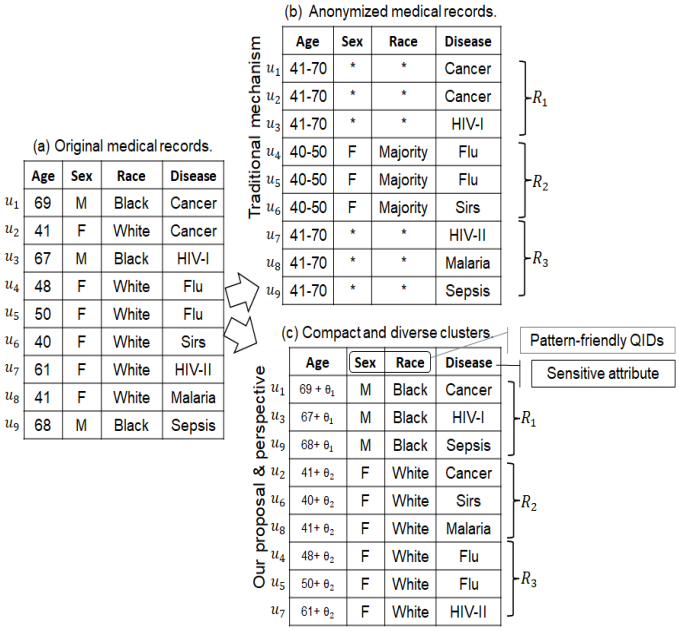


Fig. 10. Conceptual overview of the existing and the proposed anonymization scheme on personal data.

existing anonymization models have poor utility when given as input to AI applications due to extensive and unnecessary anonymization and changes. In Figure 10, we illustrate such problems with existing methods and suggest a new perspective to overcome them by taking a sample of nine medical records. As shown in Figure 10(b) (e.g., anonymization performed by existing methods), most data items are hidden, and the open attributes are overly anonymized, leading to the poor utility for general and data-mining tasks. In this paper, we suggest opening most data items, e.g., quasi-identifiers (QIDs),

without compromising record owner privacy by identifying privacy-violating and non-privacy-violating patterns, relaxing privacy budget ϵ, utilizing sensitive attribute (SA) diversity, similarity-based user grouping, and lower-level anonymization. As shown in Figure 10 (c), our proposal can retain most QID values in pure form while offering higher privacy guarantees, compared to existing methods.

Our method injects less noise in numerical QIDs than existing methods and preserves statistical information adequately. We believe that restricting changes in anonymized data can assist in retaining higher knowledge, thereby maximally improving real-world services (e.g., healthcare, reliable predictions). In the next subsections, we discuss important results from using our scheme on three datasets. To the best of our knowledge, this is a maiden approach to extracting and preserving most parts of data in their original form with strong privacy guarantees. The results and comparisons w.r.t. retaining originality in the three real-life datasets are given in Table III. From the results, it can be seen that the proposed scheme has preserved most values in their original form in all three datasets compared to previous SOTA algorithms. To the best of our knowledge, this is the maiden approach that preserves most data in its original form, thereby contributing significantly to data mining and analytical scenarios.

TABLE III
ORIGINALITY PRESERVATION: PROPOSED SCHEME VERSUS EXISTING METHODS.

D	**Proposed**	k**-anonymity [6]**	**SVD3DR [9]**	**RKA [11]**
A [22]	**80.01%**	0.0 %	0.0 %	0.0 %
B [23]	**61.88%**	0.0 %	0.0 %	0.0 %
C [24]	**40.51%**	0.0 %	0.0 %	0.0 %

E. Privacy preservation

We performed experiments by changing the degree of ν (background knowledge), and then compared the results from our scheme with existing algorithms: k-anonymity (KA) [6], RKA [9], and SVD3DR [11]. The ν used in evaluating the privacy preservation capabilities of our scheme are the true records drawn from D in partial/full form that might be available to adversaries. For example, ν can be age, sex, and race information along with some other factual information such as an individual X is part of $D^{'}$. In real scenarios, it is hard to guess which information is available to adversaries, and therefore, we chose ν randomly from D to effectively handle worst-case scenarios. We rigoriously performed a de-anonymization to find the true SA from $D^{`}$, and compared the results with the existing SOTA methods. The results and comparisons w.r.t. preserving privacy (e.g., SA disclosure) in the three real-life datasets are given in Table IV.

Referring to Table IV, the numbers such as 0.5 show the probabilistic disclosure of SA (e.g., DR) which can occur against ν. For example, in a cluster of 10 individuals, if 5 records are correctly matched based on the chosen ν, then $DR = {}^{5}/_{10} = 0.5$. In some cases, the correct matches can exist in multiple clusters, and therefore, we took the average

TABLE IV
PRIVACY PRESERVATION: PROPOSED SCHEME VERSUS EXISTING ALGORITHMS (BY VARYING PRIVACY PARAMETER (k) VALUE).

D	k	**Proposed**	k**-anonymity [6]**	**SVD3DR [9]**	**RKA [11]**
A [22]	2	0.50	0.50	0.50	0.50
	5	**0.45**	0.61	0.58	0.60
	10	**0.41**	0.81	0.69	0.71
	20	**0.35**	0.71	0.54	0.55
	30	**0.41**	0.60	0.50	0.51
	40	**0.46**	0.65	0.60	0.61
	50	**0.52**	0.68	0.63	0.66
B [23]	2	0.50	0.50	0.50	0.50
	5	**0.60**	0.82	0.75	0.80
	10	**0.51**	0.72	0.68	0.70
	20	**0.55**	0.78	0.73	0.75
	30	**0.66**	0.78	0.71	0.76
	40	**0.61**	0.80	0.76	0.78
	50	**0.62**	0.79	0.72	0.76
C [24]	2	0.50	0.50	0.50	0.50
	5	**0.35**	0.51	0.42	0.44
	10	**0.41**	0.48	0.45	0.46
	20	**0.42**	0.51	0.44	0.46
	30	**0.50**	0.58	0.55	0.56
	40	**0.52**	0.57	0.54	0.55
	50	**0.53**	0.66	0.64	0.66

of DR in respective clusters to measure the total DR. The reasons for improved results from our scheme are SA diversity preservation in each cluster, and establishing the privacy-violating partition to ensure the needed anonymization. The results in Table IV highlight how SA disclosure from using our scheme is lower in most cases than from the existing SOTA algorithms. These results verify the capabilities of our scheme concerning privacy preservation in worst-case scenarios, making it well-suited to real-life scenarios. The overall results and comparisons w.r.t. average SA disclosure by varying k values seven times and choosing ν differently in each test are given in Table V. The numbers in Table V represent the average DR computed from the seven different k values and corresponding DR given in Table IV.

TABLE V
PRIVACY PRESERVATION: PROPOSED SCHEME VERSUS EXISTING ALGORITHMS (AVAERAGE RESULTS BY VARYING k SEVEN TIMES).

D	**Proposed**	k**-anonymity [6]**	**SVD3DR [9]**	**RKA [11]**
A [22]	**0.44**	0.65	0.57	0.59
B [23]	**0.57**	0.74	0.69	0.72
C [24]	**0.46**	0.54	0.51	0.52

The SA values that were used to test the efficacy of the proposed scheme concerning privacy preservation in each dataset are income, disease, and expense categories, respectively. For example, the income information has two distinct values, $> 50K$ and $\leq 50K$ in the adult's dataset. Similarly, the disease has seventeen different values such as liver infection, liver cirrhosis, liver decomposition, liver transplant, liver carcinoma, HIV-I stage, HIV-II stage, HIV-III stage, HIV-IV stage, Alzhamir mild stage, Alzhamir moderate stage, Alzhamir severe stage, no sirs, sirs, sepsis moderate stage, sepsis severe stage, and septic shock stage in the bkseq dataset. The expense categories are of four types, $(< 5,000)$

category- I, ($< 10,000$) category-II, ($< 15,000$) category- III, and ($< 20,000$) category-IV in the care plans dataset. We performed detailed experiments to determine the disclosure level of each SA value. From the analysis and results cited above, our scheme has lower SA disclosure than previous SOTA algorithms for most k values. These results prove the main assertion of this study regarding strong privacy guarantees amid retaining most QIDs' values in their original form. These results fortify the significance of our scheme in realistic scenarios when privacy preservation is imperative.

Privacy analysis: In this section, we investigate and prove that our hybrid scheme satisfies both k-anonymity and ϵ-DP.

Theorem 1. The Laplace mechanism and k-member-based hybrid anonymity scheme satisfies ϵ-DP and k-anonymity.

Proof. Considering the parallel composition property of the DP model, an algorithm $\mathcal{M}$ with ϵ was applied to disjoint tuples, and it guarantees that whole $D^{'}$ satisfied ϵ-DP. The proposed hybrid scheme divides D into non-overlapping clusters using a k-member clustering algorithm that exploits similarities of numerical data alongside the categorical data. Due to non-overlapping (e.g., no-intersections) records in clusters, each cluster's ϵ is equivalent to overall ϵ as per the parallel composability property of the DP. There are no intersections between numeric and discrete data, and each cluster has at least k records with identical QID values in most parts, therefore, the k-anonymity criteria is also met in all clusters. Hence, the proposed scheme satisfies the ϵ-DP and k-anonymity simultaneously.

In our proposed scheme, DP was applied to numerical data and k-anonymity was applied to categorical data. The k-anonymity aims to hide the QIDs with groups, but the numerical data can break the k-anonymity. However, we experimentally prove that although the numerical values are not the same as like k-anonymity in the final $D^{'}$, the distribution of values in each cluster is aligned to that of k-anonymity. For example, in real data, if the values of age are: [33,34,39,31, 38], then the generalization interval for this data produced with k-anonymity model is likely [30-40]. Moreover, the clusters produced with DP in final $D^{'}$ have also a similar range of age values (e.g., the range of values falls between 30 to 40) or only marginally differ in some clusters, and therefore, the numeric data does not strictly violate the k-anonymity property from an analysis perspective. In some cases, there can be strict requirements of satisfying k-anonymity on both parts of data, and therefore, minor post-processing is needed in our scheme. However, it can be accomplished by utilizing generalization hierarchies of numerical QIDs and converting singular values to either interval or generalized form. In the above example, if the noised output from DP is [32,35,39,32,37], then k-anonymity can be satisfied by applying minor post-processing on the above output, and results like [30-40] or > 30 can be produced which are identical to k-anonymity. Also, if the attacker correctly figures out the numerical data of someone, he/she cannot identify the SA of that person due to categorical data anonymization and SA's diversity in each cluster. Also, the categorical data can defend against differential attacks as the frequency of values of categorical QID changes in $D^{'}$ as some records undergo anonymization and some don't. Hence, the proposed scheme can fulfill the generic properties of both models and therefore, it is fair to say that the joint use of two different kinds of methods can compliment each other in the PPDP scenario.

F. Enhancement of utility

In this subsection, we present utility results that were attained from large-scale experiments on three real-life benchmark datasets. Specifically, we present the numerical results and their comparison with the SOTA algorithms from the perspective of reduction in information loss (IL), enhancement of accuracy, and capability of preserving statistical information in anonymized data. All these criteria (s) are widely used to assess the quality of anonymized data in data-driven applications.

1) Reduction in information loss: In this subsection, we highlight the performance of our scheme in terms of reducing information loss (IL). IL is the unfortunate consequence of any anonymity scheme. However, IL can be restrained by exploiting the intrinsic characteristics of D, and applying careful anonymization. In our work, we perform only minimal and required generalization of data, and therefore, IL is restrained to the extent possible. We present the results and comparisons of using our scheme with different k values in figures 11 and 12. Figure 11 presents IL results from the non-privacy-violating partition. In this partition, most QID values are preserved as close to the original as possible by using a relaxed ϵ, lower-level anonymization, and no anonymization. IL increases with k due to an increase in records in each class. However, the proposed scheme results in significantly lower IL from most k values, compared to the existing SOTA algorithms in most cases.

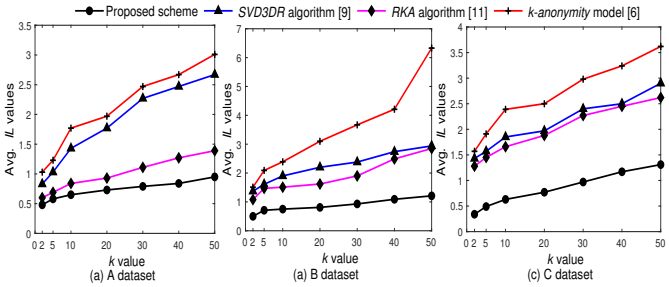


Fig. 11. IL comparisons in non-privacy-violating partition.

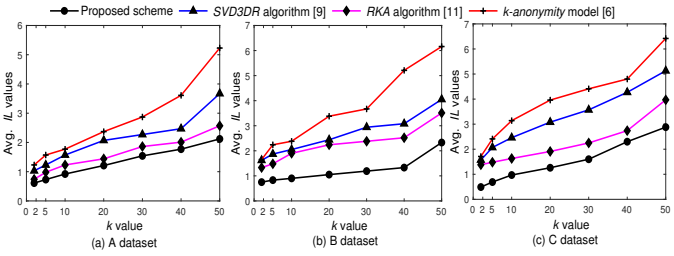


Fig. 12. IL comparisons in privacy-violating partition.

Figure 12 highlights the results in privacy-violating partitions where relatively higher anonymity is required to preserve privacy. In the proposed scheme, the diversity of an SA is considered to lower the anonymity and restrain IL.

As delineated in Figure 12, IL increases with k due to an increase in generalization and the number of records. The proposed scheme shows better performance for most k values than existing algorithms. The results given in Figures 11 and 12 prove the superiority of our scheme over prior SOTA algorithms w.r.t IL. The overall IL results and comparisons in both privacy-violating and non-privacy-violating partitions are given in Table VI.

TABLE VI
INFORMATION LOSS: PROPOSED SCHEME VERSUS EXISTING ALGORITHMS.

D	Non-Privacy violating partition				Privacy violating partition			
	Proposed	**KA [6]**	**SVD3DR [9]**	**RKA [11]**	**Proposed**	**KA [5]**	**SVD3DR [9]**	**RKA [11]**
A [22]	**0.72**	2.02	1.78	0.98	**1.27**	2.66	2.04	1.54
B [23]	**0.86**	3.33	2.16	1.85	**1.19**	3.52	2.57	2.19
C [24]	**0.81**	2.60	2.09	1.95	**1.45**	3.83	3.16	2.49

2) Enhancement of accuracy: In this subsection, we highlight the performance of our scheme in terms of enhancing accuracy. We created different variants of anonymized data and applied the RF method to get the accuracy values. The results and comparisons w.r.t. accuracy using the whole of $D^{'}$ and D are depicted in Table VII. From Table VII, it can be observed that our scheme has yielded better results in all three datasets than prior SOTA algorithms. In addition, the accuracy results of our scheme are slightly lower than the original datasets. The main reason for the higher accuracy of our scheme is due to the preservation of co-relations among QIDs and lower changes in data during anonymization. These results signify the efficacy of our scheme in analytical and data mining scenarios.

TABLE VII
ACCURACY COMPARISONS: PROPOSED SCHEME VERSUS EXISTING ALGORITHMS.

D	**Orignal data**	**Proposed**	**k-anonymity [6]**	**SVD3DR [9]**	**RKA [11]**
A [22]	0.91	0.87	0.67	0.77	0.74
B [23]	0.92	0.91	0.75	0.85	0.81
C [24]	0.82	0.78	0.64	0.71	0.68

The accuracy results and comparisons with different k values are shown in Figure 13. From the results given in Figure 13, it can be noticed that the accuracy value increases with k due to the decrease in variability in QIDs' values. Our scheme yielded higher accuracy results for most k values than previous SOTA algorithms.

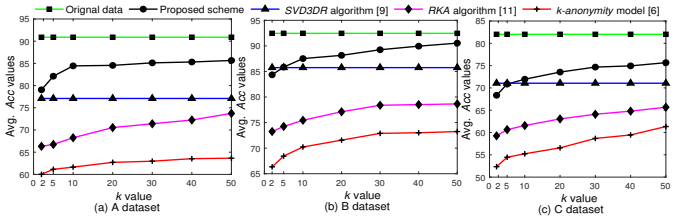
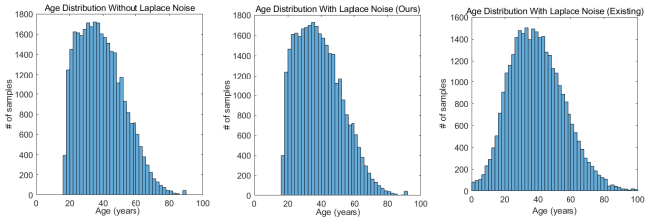


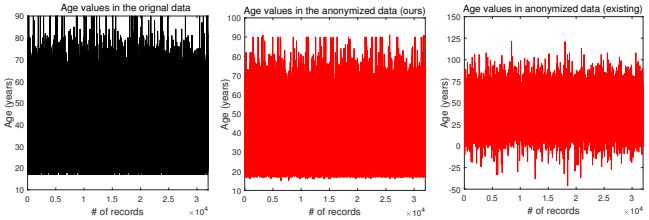
Fig. 13. Accuracy comparisons: proposed scheme versus existing algorithms.

3) Capability of preserving statistical information: In this subsection, we present the efficacy of our scheme in preserving statistical information while anonymizing numerical QIDs. Specifically, we highlight how our scheme uses a non-fixed value for ϵ in each partition, and effectively preserves statistical information (e.g., value frequency) for data miners/analysts. By preserving statistical information, knowledge discovery becomes easier, and data-specific biases can be eliminated. In addition, data with the right balance of statistical information can contribute to responsible data science (RDS)[4] and feed data-hungry applications, which are now growing and exciting areas of interest. Figure 14 compares the performance of our scheme with the existing algorithms and real data. From the results, we can see that the loss of statistical information from our scheme is marginally lower than D. In contrast, the existing algorithms had higher losses of statistical information than D even with a bounded interval [0,100]. For example, in D, no age value was less than 16, but in $D^{'}$, some values were lower than 16 (some values are even negative, as shown in Figure 15). From Figure 15, we can see that our scheme adequately preserved most age values, and there were no negative age values in the data. These results validate the effectiveness of our scheme in real-life scenario(s), especially when data of excellent quality is imperative for conducting research or validating/generating new hypotheses. The $D^{'}$ produced by our scheme is well-suited to futuristic AI data-hungry applications and RDS, where data of high quality is imperative for informed and fair decision-making. Our results align with the recent trends toward responsible use of data while preserving privacy in the AI era [25]. By preserving statistical information in $D^{'}$, data-specific biases can be eliminated in real-life scenarios when $D^{'}$ is used in training AI models.



(a) Orignal data (b) $D^{'}$ (ours) (c) $D^{'}$ (existing)

Fig. 14. Comparison of statistical information losses between D and $D^{'}$.



(a) Orignal data (b) $D^{'}$ (ours) (c) $D^{'}$ (existing)

Fig. 15. Comparison of statistical information losses between D and $D^{'}$.

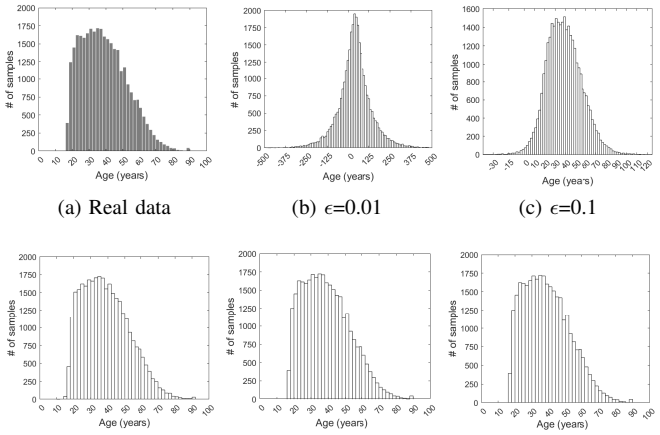
Our scheme accurately preserves statistical information using different values of ϵ in each partition. For example, it uses a relatively higher ϵ in a non-privacy-violating partition and a reasonable value of ϵ in the privacy-violating partition. However, the previous methods use a fixed value of ϵ for an entire dataset, and thereby add too much noise to the data. Applying fixed ϵ for whole datasets degrades the structure of data, and overly anonymized data has poor utility in analytical and data mining scenarios. Over-anonymized data can lead

[4]https://redasci.org/

to highly biased and unreliable results/decisions in real-life data-driven applications. In contrast, our scheme produces high-quality anonymized data that maintains an equilibrium between privacy and utility. Our scheme is most suitable for the healthcare sector where data of high quality is imperative for conducting research (or generating new hypotheses) [26]. From the results and comparisons, it can be seen that the proposed scheme strikes a balance well between individual privacy and anonymized data quality in the PPDP.

G. Effect of varying privacy budget ϵ on the performance

The small value of ϵ offers a strong privacy guarantee, but the utility loss can be very high. In some cases, the small value of ϵ can destroy the truthfulness of data, and induce fake records in the anonymized data, leading to wrong conclusions/analytics after data release. In our scheme, we applied the reasonable values of ϵ in each partition by exploiting the pattern-friendly nature of QIDs. A small value was used in the privacy-violating part, and a higher value was used in the non-privacy-violating part to strike the balance between privacy and utility. The effect of varying ϵ on change in the distribution of numerical QID in the adult's dataset is given in Figure 16. From the results, it can be seen that the value of ϵ has a great effect on data distributions. The small value of the ϵ cases a higher change in the values, and vice versa.

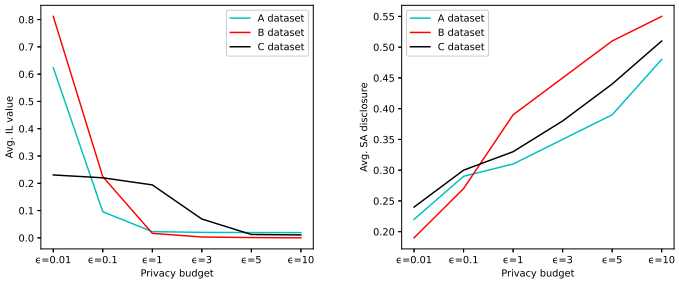


(a) Real data (b) ϵ=0.01 (c) ϵ=0.1

(d) ϵ=1 (e) ϵ=5 (f) ϵ=10

Fig. 16. Effect of varying privacy budget ϵ on QID values.

In the experimental evaluation, we used two distinct values of the ϵ for each data partition. However, the value of ϵ has a direct impact on both privacy protection and data usability. We present the effect of varying privacy budget ϵ on the proposed scheme performance, such as SA disclosure and IL in Figure 17. Referring to Figure 17, it can be noticed that IL and SA disclosure have opposite behavior with varied ϵ. The IL for numerical QIDs was determined by taking the ratio between truly preserved values in anonymized data the real data. For the higher value of ϵ (e.g., when $\epsilon \geq$ 10), the IL becomes almost zero, meaning that most values are highly similar to that of real data. Similarly, when ϵ is small (e.g., ϵ=0.01), privacy protection is higher. Furthermore, the results vary based on the characteristics of the data. For example, the B datasets yield higher IL when ϵ is small, owing to more numerical QIDs in it. In our proposed scheme, privacy is effectively safeguarded as the SA column has sufficient diversity and categorical QIDs are generalized. The IL is better as some parts of the data are anonymized rather than entire parts by exploiting the information of QIDs in a fine-grained manner.



(a) Avg. information loss (b) Avg. SA disclosure

Fig. 17. Effect of varying privacy budget ϵ on SA disclosure and IL.

H. Reduction in time complexity

In this subsection, we highlight the performance of our scheme in terms of reducing time complexity. To compute and compare time performance, we recorded the running time of our scheme on all three datasets. For fair analysis, we conducted repetitive experiments and took an average computing time after ten tests. The results and comparisons w.r.t. time complexity in the three real-life datasets are given in Table VIII.

TABLE VIII
TIME COMPLEXITY (IN s): PROPOSED SCHEME VERSUS EXISTING METHODS.

D	**Proposed**	k-anonymity [6]	SVD3DR [9]	RKA [11]
A [22]	**21.48**	142.05	90.34	129.23
B [23]	**69.53**	191.34	122.31	152.34
C [24]	**141.94**	276.12	163.35	184.19

From the results depicted in Table VIII, it can be noticed that our scheme run much faster compared to the previous SOTA algorithms. The reason for the higher time complexity of existing algorithms is the anonymization of entire sections of the data. In contrast, our scheme applies anonymity to fewer sections of data, and therefore, the time overheads are significantly small compared to previous algorithms. Hence, our scheme is also applicable in resource-constrained environments. Apart from the overall time consumption, the breakdown of the time consumed by each data processing step is provided in Figure 18. The main data processing steps in our scheme are: identifying pattern-friendly QIDs, implementing k-anonymity, and adding Laplace noise perturbation. Referring to Figure 18, steps 2 and 3 have had a relatively shorter time than step 3 because they were implemented in highly optimized libraries. However, step 2 has the highest time consumption, owing to complex operations involved in similarity and diversity consideration while making clusters. Hence, it can be concluded that implementing k-anonymity has the greatest impact on the overall method. Lastly, it is worth noting that time consumption heavily depends on the characteristics of datasets. Dataset A has less time because there exist many

records in the non-privacy violating partition, and therefore, very little anonymity was applied. In contrast, dataset B has mostly numerical QIDs due to which the noise addition time is relatively higher. The overall time for dataset C is higher because many QIDs are categorical, and higher lookups are required in generalization hierarchies during anonymization.

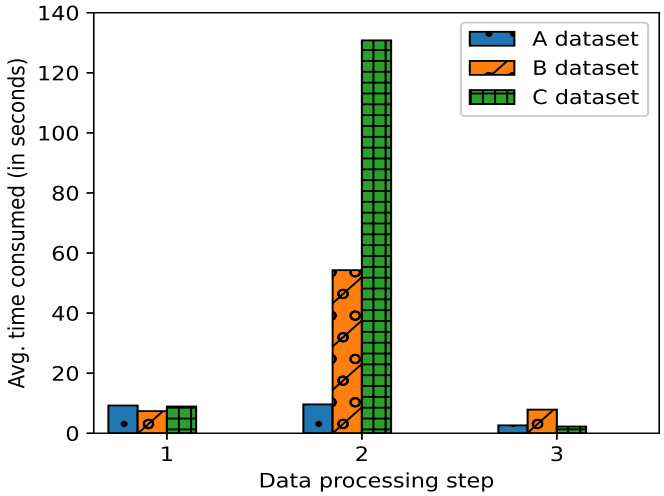


Fig. 18. Breakdown of the time consumed by each data processing step (1= Identifying pattern-friendly QIDs, 2= Implementing k-anonymity, and 3= Adding Laplace noise perturbation).

The proposed scheme is generic and can be applied to any dataset encompassed in tabular format. It can also be applied to big data scenarios (e.g., many records and many attributes are present in D) with slight modifications in relevant parameters (e.g., $ntree$, ϵ, $mtry$, k, etc.). However, while dealing with big data scenarios, our scheme may yield certain performance bottlenecks such as a rise in computing time, memory overheads, and # of operations. Also, some additional pre-processing techniques may be required to clean the data depending on the data source and quality. However, these issues can be resolved by using distributed frameworks (e.g., MapReduce), pre-computed statistics of some steps, data conversions to some easier formats (categorical $\rightarrow$ numeric, or vice versa), and data reduction strategies. Nowadays, many high-performance computing hardware (e.g., GPU/accelerators) are also available which can assist in overcoming performance bottlenecks while anonymizing big data. To lessen the data pre-processing time, some sophisticated low-code or no-code tools such as KNIME[5] can be utilized. Based on the above analysis, it is fair to say that the proposed scheme can effectively deal with big data scenarios with slight modifications.

The privacy, utility, and time complexity results cited above prove the superiority of our scheme over prior SOTA algorithms. Lastly, our scheme can ease the burden on data analysts by preserving most sections of the anonymized data in their original form.

V. Conclusion

This paper implemented a hybrid anonymity scheme for PPDP by combining differential privacy and k-anonymity, making it well-suited to futuristic AI applications and RDS, where obtaining data of excellent quality is imperative. The anonymized data not having similar functional relationships to that of real data inadvertently propagates biases in the training process of AI models. As a result, AI models can make biased or wrong decisions/predictions, leading to conflicts in society. The solution for data-specific biases is urgent to yield the intended performance with AI models. Recently, there has been a growing debate to rectify the privacy protection technologies as they may destroy important information regarding minorities from data, leading to low/no benefits for them [27]. Our scheme offers sufficient resilience against privacy attacks by identifying privacy-violating partitions and ensuring the needed anonymization in them. Different from prior methods, our scheme uses distinct values for privacy budget ϵ, thereby significantly reducing the offset in numerical QID values in real and anonymized data. The values of categorical QIDs are released, as is, in non-privacy-violating partitions, and are minimally generalized in the privacy-violating partition by exploiting SA diversity information. This is the pioneer scheme to extract and preserve most parts of data in their original form with strong privacy guarantees, making the anonymized data most suitable for data-hungry AI applications. The experiment results and comparisons from using three benchmark datasets indicate that our scheme significantly outperformed SOTA methods from the perspective of originality preservation, providing the ability to defend against privacy breaches, ensuring higher utility in the anonymized data, and significantly lowering computing overhead.

[5]https://www.knime.com/

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (2020R1A2B5B01002145).

References

[1] R. Genés-Durán, O. Esparza, J. Hernández-Serrano, F. Román-García, M. Soriano, A. Zappa, M. Serrano, S. Stahnke, B. Böhm, E. Fries *et al.*, "Data marketplaces with a free sampling service," in *2022 IEEE International Conference on Services Computing (SCC)*. IEEE, 2022, pp. 333–338.

[2] J. Wieringa, P. Kannan, X. Ma, T. Reutterer, H. Risselada, and B. Skiera, "Data analytics in a privacy-concerned world," *Journal of Business Research*, vol. 122, pp. 915–925, 2021.

[3] A. Bhardwaj and V. Kumar, "Privacy and healthcare during covid-19," in *Cybersecurity Crisis Management and Lessons Learned From the COVID-19 Pandemic*. IGI Global, 2022, pp. 82–96.

[4] L. G. Rendón, "An introduction to the principle of transparency in automated decision-making systems," in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, 2022, pp. 1245–1252.

[5] A. Majeed and S. O. Hwang, "Rectification of syntactic and semantic privacy mechanisms," *IEEE Security & Privacy*, no. 01, pp. 2–16, 2022.

[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.

[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, pp. 3–es, 2007.

[8] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*. IEEE, 2006, pp. 106–115.

[9] N. Kousika and K. Premalatha, "An improved privacy-preserving data mining technique using singular value decomposition with three-dimensional rotation data perturbation," *The Journal of Supercomputing*, vol. 77, no. 9, pp. 10 003–10 011, 2021.
[10] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.
[11] F. Song, T. Ma, Y. Tian, and M. Al-Rodhaan, "A new method of privacy protection: random k-anonymous," *IEEE Access*, vol. 7, pp. 75 434–75 445, 2019.
[12] B. Li and K. He, "Local generalization and bucketization technique for personalized privacy preservation," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 1, pp. 393–404, 2023.
[13] B. B. Mehta and U. P. Rao, "Improved l-diversity: scalable anonymization approach for privacy preserving big data publishing," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1423–1430, 2022.
[14] ——, "Privacy preserving big data publishing: a scalable k-anonymization approach using mapreduce," *Iet Software*, vol. 11, no. 5, pp. 271–276, 2017.
[15] M. Milani, Y. Huang, and F. Chiang, "Data anonymization with diversity constraints," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
[16] M. Strobel and R. Shokri, "Data privacy and trustworthy machine learning," *IEEE Security & Privacy*, no. 01, pp. 2–7, 2022.
[17] T. Yang, L. S. Cang, M. Iqbal, and D. Almakhles, "Attack risk analysis in data anonymization in internet of things," *IEEE Transactions on Computational Social Systems*, 2023.
[18] W. Huang, S. Zhou, T. Zhu, Y. Liao, C. Wu, and S. Qiu, "Improving laplace mechanism of differential privacy by personalized sampling," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2020, pp. 623–630.
[19] A. Majeed, S. Khan, and S. O. Hwang, "Towards optimization of privacy-utility trade-off using similarity and diversity based clustering," *IEEE Transactions on Emerging Topics in Computing*, 2023.
[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[21] D. Sadhya and B. Chakraborty, "Quantifying the effects of anonymization techniques over micro-databases," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 4, pp. 1979–1992, 2022.
[22] D. Newman, "Uci repository of machine learning databases, university of california, irvine," *http://www. ics. uci. edu/mlearn/MLRepository. html*, 1998.
[23] F. Amiri, N. Yazdani, A. Shakery, and A. H. Chinaei, "Hierarchical anonymization algorithms against background knowledge attack in data releasing," *Knowledge-Based Systems*, vol. 101, pp. 71–89, 2016.
[24] Z. El Ouazzani, A. Braeken, and H. El Bakkali, "Proximity measurement for hierarchical categorical attributes in big data," *Security and Communication Networks*, vol. 2021, 2021.
[25] S. Zouinina, Y. Bennani, N. Rogovschi, and A. Lyhyaoui, "A two-levels data anonymization approach," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2020, pp. 85–95.
[26] A. A. Sinaci, F. J. Núñez-Benjumea, M. Gencturk, M.-L. Jauer, T. Deserno, C. Chronaki, G. Cangioli, C. Cavero-Barca, J. M. Rodríguez-Pérez, M. M. Pérez-Pérez *et al.*, "From raw data to fair data: the fairification workflow for health research," *Methods of information in medicine*, vol. 59, no. S 01, pp. e21–e32, 2020.
[27] D. Pujol and A. Machanavajjhala, "Equity and privacy: More than just a tradeoff," *IEEE Security & Privacy*, vol. 19, no. 6, pp. 93–97, 2021.

Abdul Majeed received the Ph.D. degree in Computer Information Systems & Networks from the Korea Aerospace University, Goyang, South Korea, in 2021. He is currently working as an Assistant Professor with the Department of Computer Engineering, Gachon University, Korea. His research interests include privacy preserving data publishing, privacy-aware computing, statistical disclosure control, social network analysis and mining, and machine learning. Contact him at ab09@gachon.ac.kr.

Seong Oun Hwang received the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 2004, South Korea. He worked as a Professor with the Department of Software and Communications Engineering, Hongik University, from 2008 to 2019. He is currently working as a full Professor with the Department of Computer Engineering, Gachon University, Korea. His research interests include cryptography, cybersecurity, and artificial intelligence. Contact him at sohwang@gachon.ac.kr.