electronics

MDPI

Article

A Generic Approach towards Enhancing Utility and Privacy in Person-Specific Data Publishing Based on Attribute Usefulness and Uncertainty

Abdul Majeed * and Seong Oun Hwang *

Department of Computer Engineering, Gachon University, Seongnam 13120, Republic of Korea
* Correspondence: sohwang@gachon.ac.kr (S.O.H.); ab09@gachon.ac.kr (A.M.)

Abstract: This paper proposes a generic anonymization approach for person-specific data, which retains more information for data mining and analytical purposes while providing considerable privacy. The proposed approach takes into account the usefulness and uncertainty of attributes while anonymizing the data to significantly enhance data utility. We devised a method for determining the usefulness weight for each attribute item in a dataset, rather than manually deciding (or assuming based on domain knowledge) that a certain attribute might be more useful than another. We employed an information theory concept for measuring the uncertainty regarding sensitive attribute's value in equivalence classes to prevent unnecessary generalization of data. A flexible generalization scheme that simultaneously considers both attribute usefulness and uncertainty is suggested to anonymize person-specific data. The proposed methodology involves six steps: primitive analysis of the dataset, such as analyzing attribute availability in the data, arranging the attributes into relevant categories, and sophisticated pre-processing, computing usefulness weights of attributes, ranking users based on similarities, computing uncertainty in sensitive attributes (SAs), and flexible data generalization. Our methodology offers the advantage of retaining higher truthfulness in data without losing guarantees of privacy. Experimental analysis on two real-life benchmark datasets with varying scales, and comparisons with prior state-of-the-art methods, demonstrate the potency of our anonymization approach. Specifically, our approach yielded better performance on three metrics, namely accuracy, information loss, and disclosure risk. The accuracy and information loss were improved by restraining heavier anonymization of data, and disclosure risk was improved by preserving higher uncertainty in the SA column. Lastly, our approach is generic and can be applied to any real-world person-specific tabular datasets encompassing both demographics and SAs of individuals.

Keywords: attribute usefulness; utility; privacy; privacy-preserving data publishing; flexible data generalization; anonymization; personal data; uncertainty; privacy models; equivalence classes

Citation: Majeed, A.; Hwang, S.O. A Generic Approach towards Enhancing Utility and Privacy in Person-Specific Data Publishing Based on Attribute Usefulness and Uncertainty. *Electronics* **2023**, *12*, 1978. https://doi.org/10.3390/electronics12091978

Academic Editor: Aniello Castiglione

Received: 20 March 2023
Revised: 14 April 2023
Accepted: 20 April 2023
Published: 24 April 2023

1. Introduction

A vast amount of data/information about people is collected daily by corporations, governments, hospitals, banks, and social network (SN) service providers. Aside from demographics (age, gender, date of birth, education, marital status, etc.), the collected data often carry sensitive information, such as an individual's income, interests, hobbies, religious and political views, and health information. Publishing collected data can assist organizations in many ways, such as conducting better customer/patient analyses, creating more profitable marketing strategies, achieving strategic goals, recommending related products, boosting sales, analyzing people's behaviors, understanding the dynamics of infectious disease spread, and improving overall business performance [1,2]. However, sharing such data with data miners in their original forms may violate the individuals' privacy [3–5]. Due to privacy issues, many organizations are unwilling to publish their users/subscribers/affiliates' data for knowledge discovery with information consumers.

Generally, there are three major privacy threats that can arise from the published data analysis: unique identification of an individual (a.k.a. identity disclosure), disclosure of sensitive information (a.k.a., attribute disclosure), and disclosure of an individual's presence or absence (a.k.a. membership disclosure) [6]. To safeguard the users' privacy, data holders (e.g., hospitals) generally anonymize their user data by modifying the original values of each user's attribute before release. Privacy-preserving data publishing (PPDP) fosters data sharing with experts/researchers while ensuring individuals' privacy, on the one hand, and minimizing distortion to maintain higher utility, on the other hand [7,8]. The key techniques for safeguarding user privacy are pseudonymization, encryption, masking, anonymization, and differential privacy (DP). However, most data owners prefer anonymization for PPDP due to its algorithmic simplicity and low computing complexity. Anonymization has also been mandated by laws in some advanced countries recently [9]. In some countries, the notion of privacy is relatively weak, and it is often undervalued for national interest and economic purposes. In such circumstances, anonymizing methods that maintain higher data utility while preserving privacy are crucial. Furthermore, in some sectors, such as healthcare, sharing high-quality data is necessary for effective research and analyzing demographic connections with various diseases. However, there have been relatively few studies on methods for enhancing data utility based on attribute values [10–12]. The main motivations behind this research are as follows.

- To facilitate the analysis of released data without any constraints that DP and other recent techniques cannot provide.
- To impede higher changes in data during the conversion from their raw forms to anonymized data to enable observation of commonalities among differences, and vice versa.
- To yield consistent performances for both utility and privacy in published data analytics, and remain applicable in diverse domains for similar tasks.

In recent years, unprecedented advances in machine learning (ML) and deep learning (DL) tools, as well as information surges, have created many opportunities to extract knowledge from large and complex datasets [13–16]. These tools excel at uncovering insights about individuals from high-dimensional data using quasi-identifiers (QIs) that are often hidden or over-anonymized by existing methods. Given the significant potential of data sharing and the ever-increasing power of DL and ML tools, generating high-quality anonymized data can have profound benefits [17–20]. Therefore, there is a growing need to devise anonymizing methods that leverage the capabilities of these models to improve predictive utility while maintaining privacy. Although a wide range of solutions that integrate ML with traditional anonymizing methods to assess the quality of anonymized data have been proposed, there remains a significant lack of viable methods that consider the intrinsic and hidden properties of attributes from the original data to enhance both utility and privacy. This paper proposes and implements a generic ML-based approach to anonymization that effectively improves both utility and privacy.

The rest of the paper is organized as follows. Section 2 discusses the background and related work on anonymity methods. The system model and associated attack model are given in Section 3. Section 4 explains the proposed anonymization approach and its key steps. Section 5 discusses the experiments and results obtained from the benchmark datasets. Finally, the conclusions and future directions are presented in Section 6.

2. Background and Related Work

Data have immense potential to influence science and society. Therefore, retaining sufficient utility in anonymized data to increase their reusability is crucial [21]. In this section, we provide background information on the subject matter discussed in this study and analyze state-of-the-art studies.

2.1. Types of Attributes and Their Handling in the Anonymization Process

The raw dataset, D, contains four different types of attributes, each of which undergoes distinct treatment in the anonymization process. Detailed information on these attributes, including their definitions, examples present in D, and their handling during the anonymization process, can be found in our recent work [22].

There are two well-known mechanisms for data publishing: non-interactive and interactive [23]. We define both these mechanisms as follows.

- In the *non-interactive mechanism*, the data holder publishes the full D at once in anonymized form (e.g., after introducing some modifications to it).
- In the *interactive mechanism*, the data holder does not share the whole D in the distorted form. However, the data holder gives an interface (or text box) to data miners, with the help of the interface, data miners can run different queries on the relevant data, and gather (probably noisy) answers. DP [24] and its improved versions are mostly used in this mechanism of PPDP. Although DP provides strong privacy guarantees, the utility of the resulting dataset is often low, especially when a small ϵ is used [25,26]. Furthermore, the amount of noise injected by the DP model in less frequent parts of data is very high, leading to poor utility in data-driven applications. The utility issues and difficulty in selecting the optimal value for ϵ make DP unsuitable in scenarios where data of higher utility are required [27]. In some cases, more than one-time anonymization of the same D is required if there is some change in data (e.g., the addition and removal of tuples). To this end, the DP model iteratively increases the amount of noise in data, which means that after certain iterations, the data become completely useless [27]. Furthermore, in some cases, the DP assists in releasing aggregate information about the data, which makes knowledge discovery harder from broader perspectives. In contrast, the approach proposed in this paper offers higher utility and assists in releasing whole datasets while preserving both utility and privacy. Furthermore, it maintains utility and privacy even when repetitive anonymization of the same data is needed. The data produced with our approach can greatly contribute to observing commonalities among differences, and vice versa.

This work employs a non-interactive mechanism. The main reason for using a non-interactive mechanism is to provide complete information about the data composition, which is not possible with an interactive mechanism. The non-interactive mechanism enables researchers and data analysts to use published data for multiple analyses, such as query execution, building machine learning classifiers, rule-based analysis, pattern recognition/mining, etc. Moreover, the non-interactive mechanism is convenient for data owners as it reduces computation and communication costs [28]. In certain cases, publishing data using a non-interactive mechanism is beneficial in understanding the dynamics of a new research hypothesis/problem or a new crisis, such as the COVID-19 pandemic.

2.2. Analysis of the State-of-the-Art Anonymization Methods

2.2.1. Traditional Anonymization Methods

The k-anonymity privacy model [29] has been extensively used for the non-interactive mechanism of PPDP due to its conceptual simplicity. It preserves privacy by placing k users with identical QIs in an equivalence class (EC). Hence, the probability of identity disclosure equals $1/k$. Although k-anonymity [29] has become a benchmark method, it fails to control SA disclosure when ECs lack heterogeneity in the SA values. To resolve this problem of k-anonymity, an improved privacy model named ℓ-diversity [30] was proposed. In ℓ-diversity, each EC must contain ℓ distinct SA values. An EC satisfies the ℓ-diversity criteria if there are at least ℓ distinct SA values present in that EC. The ℓ-diversity model [30] can protect against SA disclosure to some extent. Notably, however, it can leak SA information by not considering SA value distributions. For example, an EC with 15 records can satisfy 2 diversity with two distinct SA values (e.g., 14 and 1). Hence, SA disclosure is still possible with a 93% probability for an SA value shared by 14 people in an EC. Furthermore, it cannot be applied to highly skewed data. To resolve these limitations in both models, another

promising solution named t-closeness was proposed [31]. According to this model, an EC possesses the t-closeness property if the distance between the distribution of the SA in the EC and the D is no more than a certain threshold. A table is t-close if all ECs satisfy the t-closeness property. Although it resolves the limitations of the two prior models, it still cannot overcome privacy issues due to semantic similarity between SA values (e.g., if all diseases in an EC occur in the same body part). In addition, achieving a t-close property adversely affects data quality. Thereafter, many improved versions of these three pioneer models have been proposed. Sun et al. [32] extended the k-anonymity concept with two new properties (e.g., (p, α) and p^+-sensitive k-anonymity) for controlling SA disclosure in PPDP. Similarly, Chen et al. [33] proposed the ρ-uncertainty model to control over-suppression and generalization issues for privacy protection in set-valued data.

Wong et al. [34] proposed an enhanced model based on the k-anonymity concept, named (α, k)-anonymity. It results in less data distortion and is scalable to the original size of D. Sun et al. [35] amended the ℓ-diversity model by proposing a family of enhanced (L, α)-diversity models. The authors suggested that the total weight of SA values in any QI group or EC should be at least α after modification for privacy preservation. Soria-Comas et al. [36] utilized the micro-aggregation-based method to yield t-closeness k-anonymous datasets for privacy preservation and utility enhancement. Recently, Ashkouti et al. [37] proposed a Mondrian multidimensional anonymizing method that fulfills ℓ-diversity criteria; it was devised within the Apache Spark framework (e.g., in a distributed fashion). The authors verified their method's feasibility in utilizing information loss and accuracy criteria. We refer interested readers to Zigomitros et al. [38] for more information about the methods proposed for the non-interactive mechanism of the PPDP.

2.2.2. Anonymization Methods for Enhancing Data Utility

Due to the recent advancements in ML/DL approaches and data-driven applications, a substantial number of anonymization methods to enhance classification accuracy have been proposed. Building classification models of higher accuracy by utilizing anonymized data is handy for many purposes [39]. Cagliero et al. [40] designed a classifier construction method by integrating taxonomy information to enhance structured data accuracy. Combining taxonomy information yields significant improvements in accuracy while keeping the computing efforts manageable for most of the datasets and tested classifiers. In [41], the authors developed a framework by generalizing the QIs that satisfy DP principles. That framework is a non-interactive method for publishing anonymized data, and the decision tree classifier showed better accuracy compared to other classification algorithms. Srijayanthi et al. [42] recently developed a UPA algorithm that enhances the utility of anonymized data using the clustering concept. The proposed algorithm performs feature selection and dimensionality reduction to achieve the stated goals. Chen et al. [43] devised a DP and clustering-based method to anonymize mixed data with a better balance of utility and privacy. The authors proved that k-median clustering combined with the DP can create anonymized data with reduced IL and time than existing SOTA methods. Jha et al. [44] devised a k-anonymity-like model called, z-anonymity for anonymizing stream data. The proposed model can be applied to similar problems, and it yielded lower IL when computed using the entropy concept. Li et al. [45] developed a bucketization and local generalization-based anonymization method to protect identity and SA disclosure in PPDP. The utilities of the resulting data were analyzed using aggregate query answering and NCP metrics. Chu et al. [46] developed a DP-based method, known as the SFLA–Kohonen Network for PPDP scenarios. The proposed method has the ability to reduce privacy risks without compromising g the usability of anonymous data.

Sun et al. [47] developed a DP-based method for preserving utility and privacy in trajectory data sharing. The proposed solution uses synthetic data to create anonymized data, and it has significantly enhanced the utility of anonymized data. Nóbrega et al. [48] developed a transfer learning-based approach for the privacy-preserving record linkage problem. The proposed approach determines a suitable threshold that can be used to

prevent record linkage in published as well as public datasets. Amiri et al. [49] recently developed a UHRA algorithm for enhancing privacy and utility in data-sharing scenarios. The proposed algorithm safeguards privacy breaches against background knowledge attacks while sustaining data utility and privacy. Chen et al. [50] devised a framework to measure the level of privacy and usefulness offered by the anonymized datasets. Xia et al. [51] developed a clustering and DP-based model for reducing IL in data-releasing scenarios. The proposed approach groups identical records into the same EC to lower IL. Han et al. [52] devised a new anonymity method considering data availability for data mining purposes. Their proposed weighted full-domain anonymization (WFDA) algorithm adaptively anonymizes data (i.e., generalizes QIs of high weights to lower levels, and vice versa). Last et al. [53] presented an anonymization algorithm by coupling non-homogeneous generalization with SA value distribution. The proposed method yields higher predictive utility compared to prior methods. Furthermore, some classifier-based anonymity methods have been discussed in the literature [54–56]. Eyupoglu et al. [57] discussed a new anonymity algorithm by utilizing chaos and perturbation concepts. The proposed chaos and perturbation algorithm (CPA) yields better results w.r.t. probabilistic anonymity, Kullback—Leibler divergence, accuracy, and the F-measure. Ye and Chen [58] proposed an algorithm to reduce information loss (IL) in the PPDP. Kousika et al. [59] developed an ML-powered anonymization method in order to improve the utility of anonymized data for training classifiers. The proposed approach integrates SVD and RDP methods to perturb the data. Detailed comparisons (e.g., the main focus of each study and possible attacks) of the recent SOTA utility-aware anonymization methods are given in Table 1. However, there are three major problems with the aforementioned techniques.

- The existing literature gives equal importance to each QI from a utility point of view or determines useful QIs from a dataset solely based on assumptions/domain knowledge without analyzing the underlying values [10,11], which results in a higher amount of utility loss and privacy breaches.
- The existing studies enforce constraints (i.e., strict parameters, ℓ, t, ρ, α, etc.) with predetermined values on the values of SA in each EC. However, this can lead to inconsistent generalization intervals due to excessive shuffling of records [30–33]. In skewed data, enforcement of such constraints is not practically possible due to less heterogeneity in some SA values.
- There is a lack of methods that can simultaneously leverage attribute usefulness and uncertainty information to control unnecessary generalization, in order to enhance both utility and privacy in anonymized datasets.

Table 1. Detailed comparisons of the recent SOTA utility-aware anonymization methods.

Study	Main Focus of Study		Possible Attacks				
	P	**U**	**H**	**S**	**SS**	**BK**	**SAI**
Li et al. [39]	×	✓	✓	✓	✓	○	✓
Cagliero et al. [40]	×	✓	✓	✓	×	✓	✓
Zaman et al. [41]	✓	✓	-	-	-	✓	○
Srijayanthi et al.[42]	×	✓	✓	✓	✓	✓	✓
Chen et al. [43]	×	✓	×	○	○	✓	×
Jha et al. [44]	✓	×	✓	○	✓	×	✓
Li et al. [45]	✓	×	✓	×	×	○	×
Chu et al. [46]	×	×	✓	✓	×	○	✓
Sun et al. [47]	×	✓	-	✓	✓	✓	○
Nóbrega et al. [48]	✓	✓	×	✓	✓	×	✓
Amiri et al. [49]	✓	✓	×	×	✓	×	✓
Chen et al. [50]	✓	×	×	✓	✓	✓	✓

Table 1. *Cont.*

Study	Main Focus of Study		Possible Attacks				
	P	**U**	**H**	**S**	**SS**	**BK**	**SAI**
Xia et al. [51]	✓	✓	×	✓	✓	✓	○
Han et al. [52]	×	✓	✓	✓	✓	○	✓
Last et al. [53]	×	✓	✓	○	✓	○	✓
Fong et al. [54]	×	✓	✓	○	✓	×	✓
Lin et al.[55]	×	✓	✓	✓	✓	×	✓
Park et al. [56]	✓	×	○	○	○	×	○
Eyupoglu et al. [57]	✓	✓	✓	✓	✓	×	×
Ye et al. [58]	✓	×	✓	✓	✓	○	✓
Kousika et al. [59]	×	✓	×	✓	✓	✓	✓
Our approach	✓	✓	×	×	○	○	×

Abbreviations: P (Privacy), U (Utility), H (Homogeneity attack), S (Skewness attack), SS (Semantic similarity attack), BK (Background knowledge attack), SAI (SA inference attack). **Key**: ✓⇒ considered/possible, ×⇒ not-considered/not-possible, ○ ⇒ partially possible, and - → not applicable.

2.3. Major Contributions of This Work

In this paper, we solve the aforementioned problems of existing techniques by proposing attribute usefulness and an uncertainty-aware flexible anonymization approach. The major contributions of this paper are summarized as follows.

- We devised an RF-based method with a QI value shuffling strategy to identify the useful attributes from original data in an automated way that have minimal impact on individuals' privacy, instead of manually deciding or assuming that certain attributes might be more useful than others.
- We employed the information theory concept (i.e., entropy) for computing the uncertainty of SA values in ECs to limit privacy breaches in low-uncertainty ECs and to increase utility in high-uncertainty ECs.
- We propose a flexible data generalization method for anonymizing data that takes into account the usefulness weights of the QIs as well as the uncertainty of the SAs, which enhances two competing goals.
- The proposed approach can be used to produce anonymized versions of any dataset, whether balanced (i.e., the SA value distribution is uniform) or imbalanced (i.e., the SA value distribution is skewed).
- This is the first generic approach toward enhancing both utility and privacy by retaining (in the anonymized data) highly useful QI values that are as close as possible to the original values.

Extensive experiments were conducted on real-world and benchmark datasets, encompassing the QIs and SA of individuals under different conditions to verify the efficacy of our approach. The experimental analysis indicates significant improvements in both goals compared to prior SOTA methods and models. The proposed approach significantly controls heavier changes in QI values while converting raw data into anonymized data.

3. System Model

In this paper, we consider a common scenario for data publishing with multiple actors involved. Figure 1 provides an overview of data publishing for information consumers. The primary actors in this scenario include the user (also known as the record owner), which is an entity that is associated with one or more records, the data holder (also known as the owner), which is an organization or person that holds the users' data, the data publisher (also known as the releaser), which is a person or organization that publishes the data, the data analyst (also known as the data recipient), which is an entity that has access to the published data to extract useful knowledge from it, and the adversary (also known as the attacker), which is an entity whose goal is to compromise the user's privacy from the released data. Our goal is to devise a secure approach for data publishing that

simultaneously achieves two competing goals, i.e., (i) the analysts can extract maximum knowledge from the data published by the data owners, (ii) an adversary should not be able to compromise any user's privacy even if he/she has access to the abundance of auxiliary data (or background knowledge).

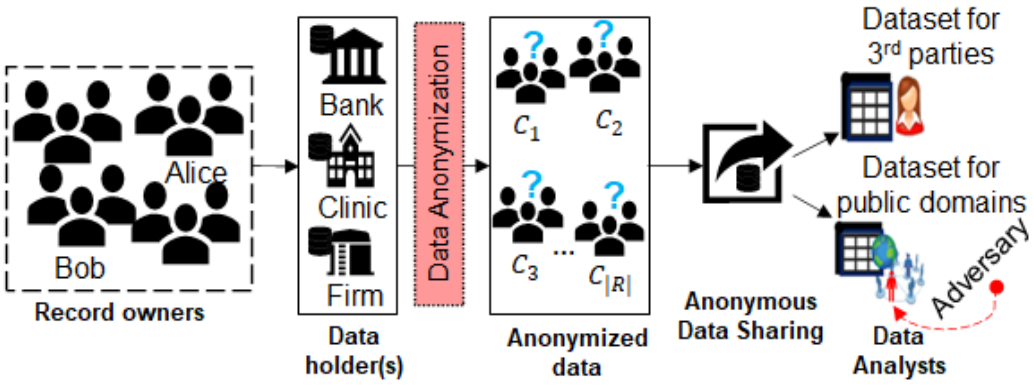

Figure 1. Overview of the system model considered in this study.

We assume that the dataset denoted by D, has already been collected from relevant users, and each row in D represents a real-world person with its public (QI) and sensitive (SA) information, respectively. Any D encompassing QI and a single SA can be processed with our model.

3.1. Attack Model

In this work, we assume that data owners/users, data holders, and data publishers are honest. They perform desired actions only and assist in accomplishing the task of data sharing. However, data analysts are honest but curious (e.g., they can behave similar to an adversary), and try to compromise the privacy of individuals. Although we remove directly identifiable information of all kinds, QIs can be learned/linked from various sources to re-identify people uniquely. Hence, the proposed method is vulnerable to identity and corresponding SA disclosures because adversaries may:

- Already know a part of the released QIs of an individual and attempt to figure out the rest of the QIs. For example, an adversary may know the age and gender value of an individual and try to identify the individual's zip code.
- Already know the whole record (e.g., all QIs) of the individual and that the respective individual is part of the released data with higher probability. Based on this information, he/she tries to obtain sensitive information from the released data concerning that individual. For example, the released data can encompass sensitive data (e.g., disease contracted, monthly income, etc.) about individuals. If the adversary can somehow identify/link the user's QIs correctly, he/she can also know the SA related to that user.

In this regard, we aim to safeguard users' privacy against these contemporary privacy threats (i.e., inference of identity and associated SA) that can emerge during published data analysis.

3.2. Design Goals

This paper aims to design a generic, utility-enhanced, privacy-preserved, and practical anonymization approach. Specifically, our approach is designed to achieve the following two major goals in PPDP.

Utility: Ensures that data analysts can perform unconstrained analysis on released data, and can extract the enclosed knowledge up to the maximum extent to improve decision-making. The proposed approach allows for both general (e.g., frequency analysis, commonalities, differences, patterns, etc.) and special (e.g., analytical and data mining

tasks) purpose analyses. Specifically, it lowers information loss and maximizes classification accuracy.

Privacy: Guarantees that an adversary with an abundance of auxiliary information cannot link/match an individual's QIs and infer his/her SA with a higher probability. It ensures that when an adversary attempts to link a user between released and auxiliary data, there exists a linkage of any record to multiple SA values. Specifically, it lowers the probabilistic disclosure of identity and corresponding SA value.

3.3. Problem Formulation

The main problem that we seek to solve in this paper is formally stated in Problem 1.

Problem 1. *Given a relational dataset D, with user attributes including age, gender, name, race, weight, country of origin, and income/disease, along with a privacy parameter k, how can we create an anonymized dataset $D^{'}$ that satisfies the k-anonymity criterion and achieves the two competing goals stated in Section 3.2. The objective model of $D^{'}$ is formulated as*

$$\begin{cases} D^{'} = [Q, S] = (\{QI_1, QI_2, QI_3, \cdots, QI_n\}, s_i) \\ Minimize \quad f_1(D^{'}) = information\,loss(D^{'}) \\ Maximize \quad f_2(D^{'}) = accuracy(D^{'}) \\ Minimize \quad f_3(D^{'}) = probabilistic\,disclosure(D^{'}) \\ Minimize \quad f_4(D^{'}) = homogeneous\,attack(D^{'}) \end{cases}$$

4. Attribute Usefulness and Uncertainty-Aware Anonymization Approach

To effectively address the key problems cited in Section 2.2.2, we propose a novel anonymization approach that exploits useful knowledge (from raw data) to anonymize them without compromising both utility and privacy. Table 2 presents the main notations used in the proposed anonymization approach.

Table 2. Main notations and their descriptions.

Notation	Description
D	Original data (i.e., before anonymization)
$D^{'}$	Anonymized data (i.e., after anonymization)
N	Number (#) of records/tuples in a D, where $N = \|D\|$
u_i	ith user in a D
Q	Quasi-identifiers' set
S	Set (a.k.a domain) of SA values
Ω_{QI_p}	Usefulness weight of pth QI
ζ	Set of usefulness weights of the QIs
$Sim(u_i, u_j)$	Similarity between two user i and j
k	Privacy parameter
τ	Total number of ECs to be made from D
X	Matrix of highly similar users
R	Set of EC made based on k
C_i	ith equivalence class with at least k users
$U(C_i)$	Uncertainty value of the SA in ith EC
T_U	Threshold value for analyzing U
U	set of the uncertainty values of ECs
H_{QI_i}	Generalization hierarchy of an ith QI

The main motivation for this novel anonymization approach is to treat QIs based on their impact on data utility, which has been largely unexplored in previous studies. For instance, in decision-making based on two QIs (gender and age), gender has a greater impact on whether a person will participate in a computer science (CS) class or not because females are often less likely to take CS-related courses. Similarly, in terms of privacy,

zip code has a greater predictive capacity to locate an individual accurately than race or birthplace. Hence, these real-world examples suggest that the weight of each QI varies in terms of both utility and privacy. To effectively preserve utility for information seekers and data miners, highly useful QI values should retain their original forms as much as possible. To this end, our proposed approach mainly focuses on enhancing utility and privacy by automatically exploiting the underlying values of each QI and controlling significant changes in QI values. Figure 2 highlights the workflow of our proposed approach.

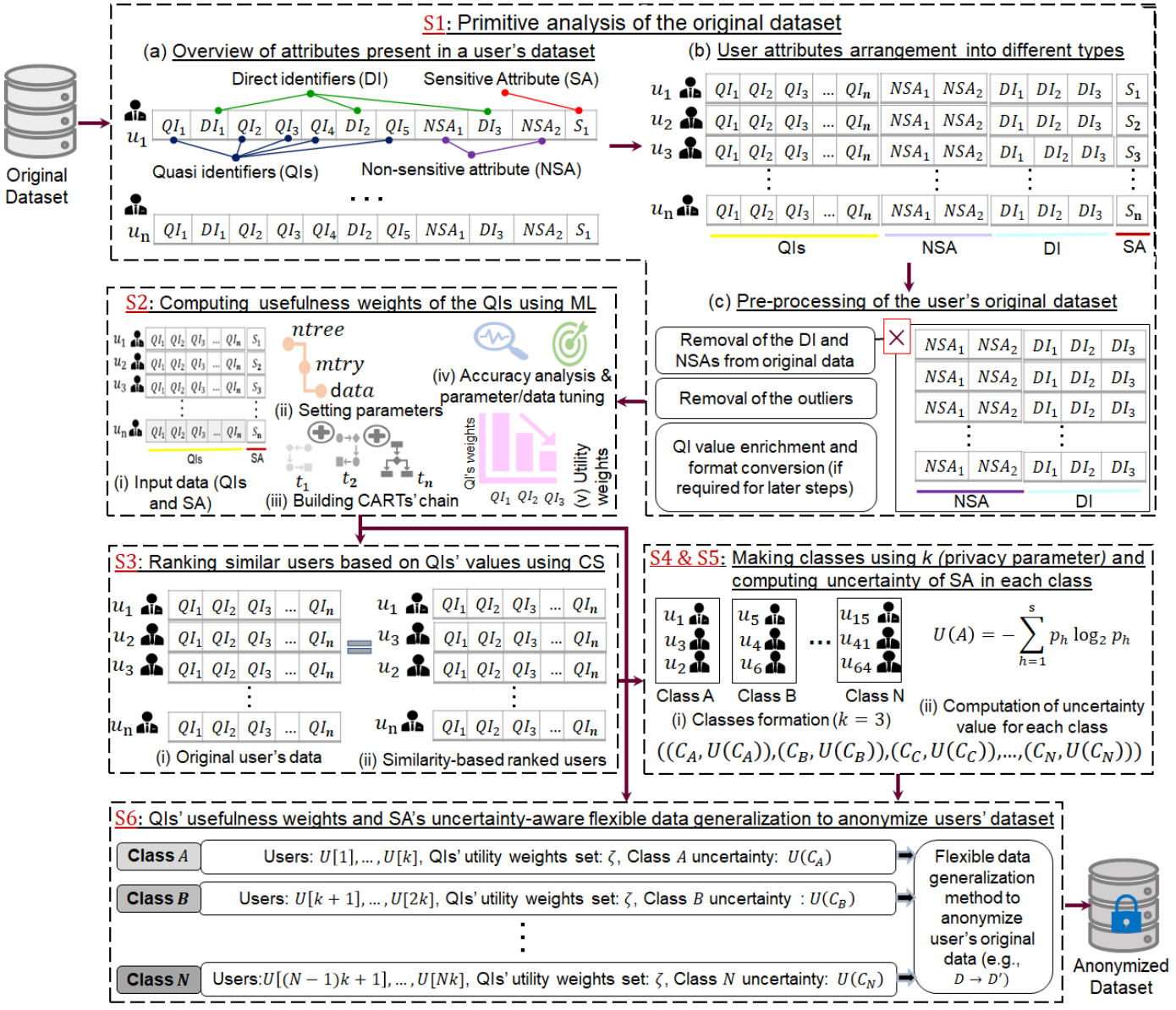

Figure 2. Conceptual overview of the proposed anonymization approach (S1−S6 denotes key steps).

To produce the anonymous version, $D^{'}$, of any original user dataset, D, containing N records (i.e., users) with n QIs and a single SA, S, we propose six principal concepts: (i) primitive analysis of the original users' dataset, D; (ii) computing usefulness weights for the QIs using an ML approach (random forest); (iii) ranking similar records via their QI values by employing cosine similarity; (iv) making equivalence classes from a similarly ranked users matrix based on the parameter of privacy (e.g., k), (v) computing the uncertainty of each EC w.r.t. SA values; and (vi) attain flexible data generalization by simultaneously considering both the usefulness weights and the uncertainty. All concepts are interlinked and help in transforming D into $D^{'}$. Concise details of each concept are presented below.

4.1. Primitive Analysis of the Original Dataset

D contains sensitive and public information about each user, u_i, depending on the phenomena for which data are collected. In a hospital scenario, public information can be the patient's name, age, gender, place of residence, etc., but sensitive data can be his/her disease information. In this paper, we assume that data about a specific purpose has already been collected from the relevant users. We performed a primitive analysis of D before its anonymization with the help of the three steps explained below.

4.1.1. Analysis of the Attributes Present in a Dataset

After obtaining D from the data owner, O_i, we analyze D's structure. This includes understanding the types of data (categorical, numerical, or hybrid), and which types of attributes are present in D because, in some cases, non-sensitive attributes (NSAs) are not collected from users. Furthermore, the range of values is checked, and dimensions are analyzed, including the number of rows and columns. If they are missing, labels are assigned to the rows and columns of D. An overview of the label assignment is depicted in Equation (1).

$$D_{Users,Attributes} = \begin{pmatrix} u_i & QI_1 & QI_2 & QI_3 \cdots & QI_n & S \\ u_1 & v_{QI_1} & v_{QI_2} & v_{QI_3} \cdots & v_{QI_n} & v_{S_1} \\ u_2 & v_{QI_1} & v_{QI_2} & v_{QI_3} \cdots & v_{QI_n} & v_{S_2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ u_N & v_{QI_1} & v_{QI_2} & v_{QI_3} \cdots & v_{QI_n} & v_{S_2} \end{pmatrix} = \begin{pmatrix} u_i & QI_1 = age & QI_2 = gender & QI_3 = eyecolour & S = salary \\ 1 & 29 & Male & Black \cdots & < 50K \\ 2 & 40 & Female & Black \cdots & > 50K \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 3000 & 34 & Male & Black \cdots & < 50K \end{pmatrix} \quad (1)$$

In Equation (1), each row represents an individual, u_i, where $i = 1, 2, 3, \cdots, N$, and each column represents the value of the particular attribute. Because the attributes can be randomly placed in D, attribute arrangement is required to simplify the subsequent operations and lower the computational complexity of the anonymization process.

4.1.2. Arrangement of Attributes into Appropriate Categories

After the label assignments, attributes are arranged into their respective categories. For example, all QIs will be placed under the QI category, and so on. As shown in Figure 2, attribute handling becomes relatively easy when attributes are arranged in their respective categories. We apply pre-processing to the arranged attributes to obtain complete information about each individual.

4.1.3. Pre-Processing of the D to Yield Informative Analysis

Data pre-processing has become an integral component in data-driven applications [60]. It is extremely useful and also necessary in an anonymization mechanism to increase data utility afterward. It reduces computing power significantly by removing redundant data and unnecessary features/attributes. In the pre-processing step, the proposed approach removes the NSAs and the DIs as per the standard policy of the PPDP. Subsequently, D contains only the set of QIs, Q, and the SA, S, as formally expressed in Equation (2).

$$D = \{Q, S\} = (\{QI_1, QI_2, QI_3, \cdots, QI_n\}, s_i) \quad (2)$$

Later, we find and remove outliers, such as unrealistic values for certain attributes (e.g., an age value of 500 years instead of 50), by visually inspecting the logical range of each attribute and performing min–max analysis. Outlier detection and removal significantly contribute to accurate data analytics. In some cases, we perform format conversion and enrichment to obtain desired statistics before actual data anonymization. For instance, we represent categorical values of the quasi-identifiers (QIs) with numbers to compute similarity/distance between users. Additionally, missing values are effectively handled by considering the nature of the attributes. For example, we impute missing values in

numerical attributes with the mean of the particular column, while we substitute missing values in categorical columns with less representative values to increase diversity in data. We also remove redundant records that are located next to each other and have identical values in a row. A comprehensive overview of the assumptions and limitations of the pre-processing process is demonstrated in Figure 3. Overall, pre-processing allows us to achieve a truthful and error-free *D* for further operations.

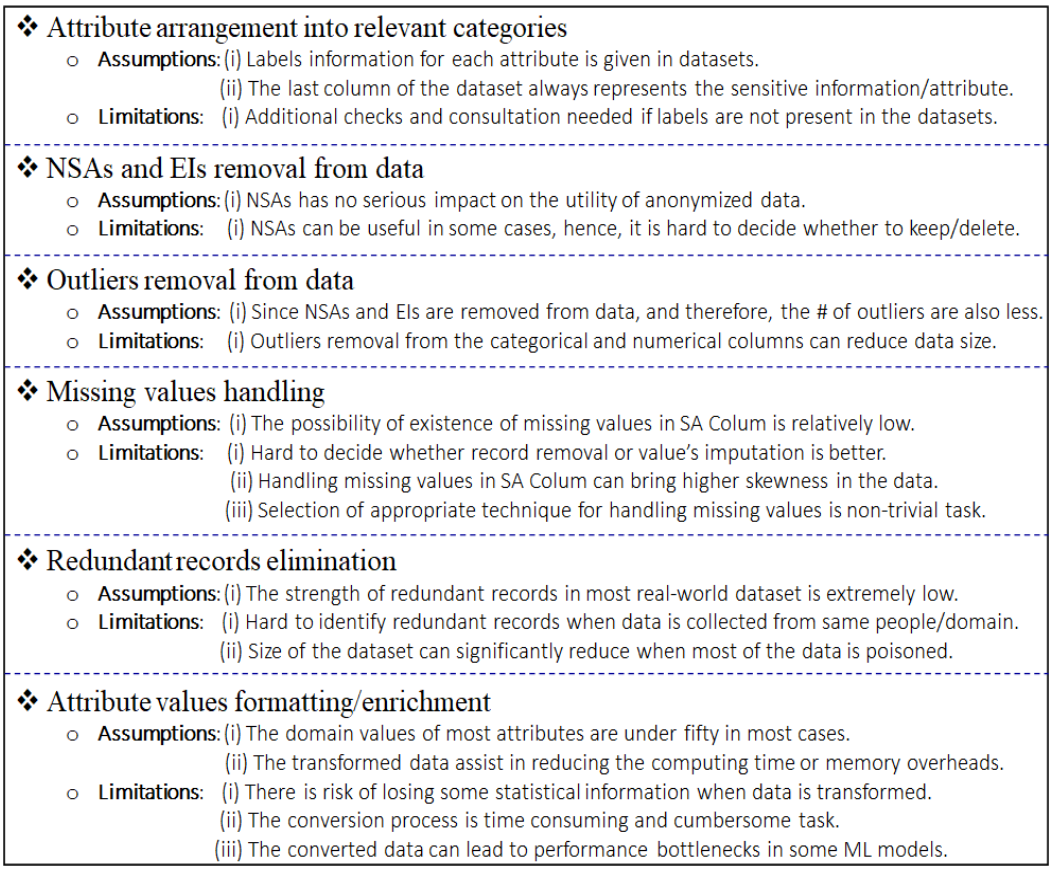

Figure 3. Overview of assumptions and limitations of the pre-processing step.

4.2. Computing Usefulness Weights of QIs in a Dataset Using Machine Learning

Many studies suggest that each QI should be treated unequally from the privacy and utility points of view to effectively address the privacy—utility trade-off (PUT) [57,61]. On the one hand, attribute-based similarities can be exploited to identify unique people from published data [61]. On the other hand, attributes-based profiling of users can assist in understanding research problems more easily, such as modeling epidemic-disease dynamics [62]. Therefore, considering the importance of unequal treatment of each QI present in *D*, we developed a method for quantifying the usefulness weights of the QIs by leveraging the ML approach, called random forest (RF) [63]. RF [63] is a versatile ML approach that belongs to the ensemble learning techniques. We selected RF to quantify QI usefulness weights because it is highly accurate compared to other currently available ML algorithms. It considers the interaction between the QIs and yields higher accuracy values. Furthermore, it scales well with the large number of records and QIs in *D*. It gives reliable results by building a chain of trees, rather than relying on a single tree's results (i.e., a decision tree).

In addition, the unique features of RF, such as the ability to handle outliers, missing values, and large records, have made it a very popular ML method. It gives accuracy/error-rate results by accumulating each tree's result via majority voting or summing of errors. Hence, results are more reliable and accurate. It is worth noting that a gradient-boosting algorithm can also be used for a similar task (e.g., identifying useful QIs from data). How-

ever, the selection of optimal hyperparameters is very challenging and computationally expensive. Furthermore, gradient boosting often experiences the problem of overfitting when unoptimized hyperparameters are used. Rigorous parameter tuning and extensive cross-validation increases are needed while experimenting with gradient boosting, leading to higher computing costs [64]. In contrast, it has fewer parameters, which can be tuned flexibly in considering the problem size. In addition, the utilization of information–theocratic concepts such as the Gini index/Gini impurity in the tree construction process makes it more suitable for data-driven applications. In this paper, we employ RF to identify and rank the most useful QIs from D to effectively enhance the utility of $D^{'}$. A descriptive block diagram of the procedures employed to compute usefulness weights is given in Figure 4.

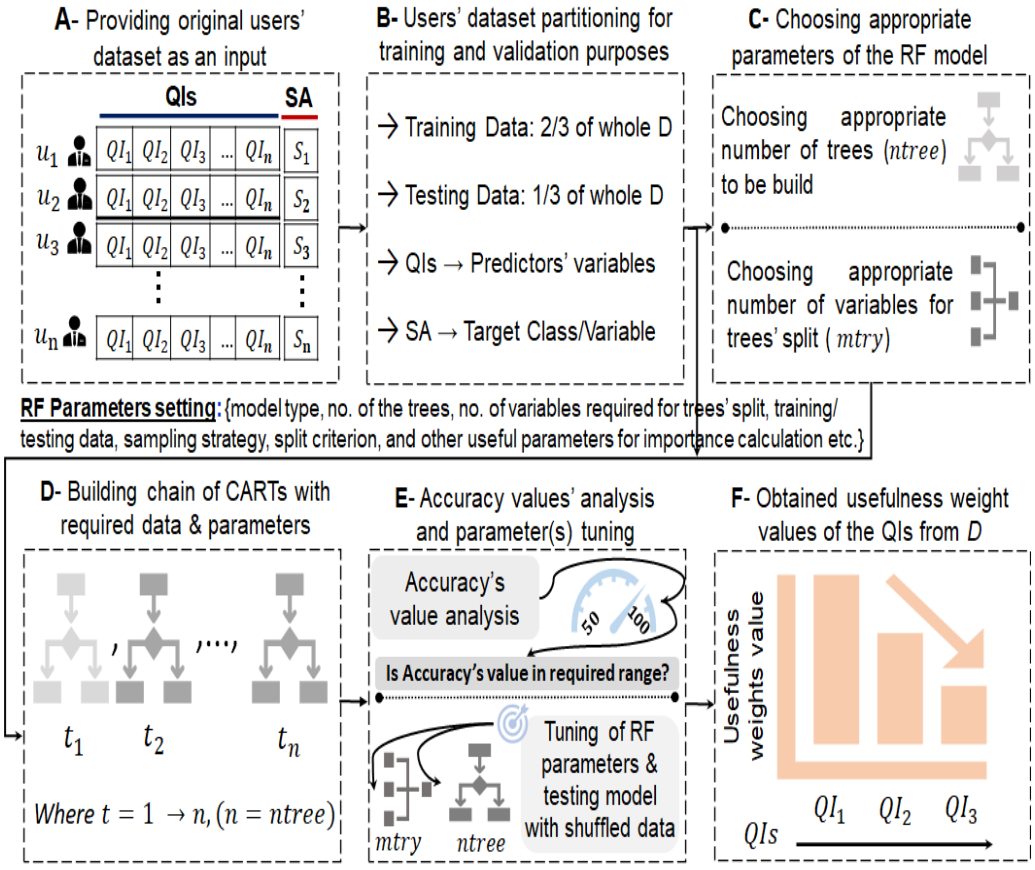

Figure 4. Block diagram of procedures for computing QI usefulness weights.

In some cases, identifying useful attributes from the original data in an automated way may be challenging. For example, the usefulness of certain attributes may only become apparent after processing the raw data. In such cases, data owners may need to rely on domain experts or manually select attributes that are deemed more useful. However, this approach is only feasible when the dataset is small or when data owners seek assistance from experts. Moreover, randomly selecting a useful attribute based solely on its label may lead to privacy issues, especially when the attribute has many distinct values. In this paper, we propose an automated method that leverages the RF model to pinpoint useful QIs from the data that no longer pose serious threats to individual privacy. Specifically, our RF-based method rigorously analyzes each QI value in relation to SAs and ranks QIs based on their utility. This process of identifying useful QIs is crucial because only some parts of a QI can pose risks, while most parts exhibit general patterns that can be released as is or can be minimally generalized for information consumers [20]. In this regard, our work offers a new perspective to the database and privacy community on how to identify useful attributes based on their values in an automated (rather than manual) way.

Although our work is unique in terms of identifying useful attributes from D that no longer pose threats to an individual's privacy, the weighting mechanism can also be controlled by the individual data owners. Notable developments in this line of work include the concept of personal privacy, facilitated by a privacy language from the SPECIAL project [65], the layered privacy language (LPL) [66], and privacy preference policies such as Contra [67]. These developments provide users with the ability to define the processes for how their data can be processed, used, and eventually removed from the system. They also ensure compliance with privacy laws such as the General Data Protection Regulation (GDPR), which restricts the manipulation of personal data. Some of these technical procedures, such as Contra, can be utilized in resource-constrained environments to meet legal and technical privacy requirements. In most of these developments, users specify their consent and other privacy preferences, and the system respects users' choices while processing their data. Most of these solutions were developed to ease the difficulties of the industry with GDPR compliance and to ensure respectful treatment/handling of personal information. In contrast, this work assumes that users do not specify any privacy preferences/consent while providing their data to data owners. However, the data owners ensure the protection of personal information while outsourcing data by applying a strong anonymization mechanism (removing directly identifiable information) to safeguard individuals' privacy while providing considerable utility for data miners/analysts. Our work enables the secondary use of person-specific data while lowering the possibility of privacy breaches.

To compute the usefulness weights of QIs, we employ six fundamental concepts labeled A to F in Figure 4. The input for computing the usefulness weights is dataset D, which contains both QIs and SAs. Next, two partitions of D (training and testing data) are created for training and validation, respectively. It is worth noting that most real-world datasets are noisy, skewed, and can have multiple vulnerabilities (e.g., outliers, small size, incomplete tuples), leading to biased results when fed into ML models. In the experiments, we used a fair sampling strategy to control bias in data. Specifically, the samples were drawn with a relatively higher sampling fraction ratio (0.80) to have tuples from each SA category value. Furthermore, at the pre-processing stage, data were cleaned using multiple techniques to control biases in data. Moreover, the training and test data were inspected from the perspective of balancedness, meaning that the distribution of each SA value was balanced with respect to real data. QIs are regarded as predictor variables, and the SA is the target class/variable. Subsequently, the RF parameter values (*ntree*, denoted with t, where $t = 1\,upto\,ntree$, and *mtry*) were chosen by considering the data sizes, and chains of trees were built from the training data samples. The choice of optimal values of the parameter is imperative to prevent imbalanced learning, leading to biased results. To prevent imbalanced learning, the grid search method was used to determine the optimal values of the parameter. Furthermore, rigorous experiments and sensitivity analyses were performed by varying each parameter's value to find the optimal values. By choosing the feasible values of each parameter, the bias was restrained in the training process. Once all trees were constructed, the accuracy value a_{int} was analyzed and validated with the testing data. If a_{int} was very low, then RF's parameters were tuned to bring a_{int} into the required range. During this process, the out-of-bag (OOB) error, represented as σ_{int}, was reduced. These values (a_{int} and σ_{int}) are recorded as reference values. Once, the reference values were obtained, the QI values were shuffled, and the RF model was built again with the shuffled data. By doing so, the correlation between the actual SA values and the QI values was demolished, and when the model was rebuilt, the accuracy values decreased if a QI strongly related to an SA. In contrast, if the accuracy changed slightly or increased, it meant that particular QI was more useful with a large number of identical values. We recorded the values for accuracy and OOB errors after the QI values were shuffled into a_f and σ_f, respectively. We took the difference, χ, of errors both before and after permutation to carry out further steps to compute QI usefulness weights, as expressed in Equation (3).

$$\chi = \sigma_{int} - \sigma_f \tag{3}$$

where σ_{int} is the OOB error before permuting the QI values, and σ_f is after the permutation. We compute the cumulative errors χ_t from all trees t of the RF method by using Equation (4).

$$\chi_t = \sum_{t=1}^{ntree} \chi_t \tag{4}$$

We compute the average, variance, and standard deviation of these errors using Equations (5)–(7).

$$x = \frac{\chi_t}{t} = \frac{\sum_{t=1}^{ntree} \chi_t}{t}, \; where \, t = |ntree| \tag{5}$$

$$s^2 = \frac{1}{t-1} \sum_{t=1}^{ntree} (\chi_t - x)^2 \tag{6}$$

$$s = \sqrt{\frac{1}{t-1} \sum_{t=1}^{ntree} (\chi_t - x)^2} \tag{7}$$

After the above calculations, we find the usefulness weight of the QI, QI_p, using Equation (8).

$$\Omega_{QI_p} = \frac{x}{s} \tag{8}$$

By using Equation (8), the weight values of all QIs are computed, and QIs are ranked considering their usefulness weights. Next, we normalize the weights, so the sum of all weights is equal to 100. By employing the method stated above, we can find and categorize QIs as highly useful, of medium value, or less useful. These statistics are verified from the actual D and are utilized in the anonymization process. The output of the method described above is the set of ranked QI-usefulness weights, ζ, (in order: high, medium, and low) as expressed in Equation (9).

$$\zeta = \{(\Omega_{QI_1}, \Omega_{QI_2}, \cdots, \Omega_{QI_h}), (\Omega_{QI_1}, \Omega_{QI_2}, \cdots, \Omega_{QI_m}), (\Omega_{QI_1}, \Omega_{QI_2}, \cdots, \Omega_{QI_l})\} \tag{9}$$

Retaining highly useful QI values as close to the original as possible during data generalization significantly contributes to data reusability and meets data miners' needs.

4.3. Ranking Similar Users to Lower Generalization Intervals

To maintain domain consistency in the QI values during data generalization, users in each EC must be highly alike. By doing so, distortion in $D^{'}$ can be controlled dramatically. In this work, we compute the similarity values between users by leveraging their QI values with the help of cosine similarity [68]. It is a simple but very effective measure for computing homophily among users in diverse domains [69]. The mathematical equation employed to measure *Sim* among two users, u_1 and u_2, is given in Equation (10).

$$Sim(u_1, u_2) = \frac{\sum_{i=1}^{n} u_{1_{(i)}} \times u_{2_{(i)}}}{\sqrt{\sum_{i=1}^{n} u_{1_{(i)}}^2} \times \sqrt{\sum_{i=1}^{n} u_{2_{(i)}}^2}} \tag{10}$$

In Equation (10), u_1 and u_2 are two distinct users, i denotes QIs, and n denotes the total number of QIs in D. Furthermore, $Sim(u_1, u_2) \in [0, 1]$. If the value of *Sim* is 0 (e.g., *ii* in Equation (11)), it implies that nothing is common between the two users (i.e., they are highly dissimilar). In contrast, a value of 1 for *ii* in Equation (11) (*i* in Equation (11)) implies the two users are the same (i.e., all QIs have identical values). Apart from these two cases,

Sim values range between 0 and 1 (*iii* in Equation (11)). We express the expected value of *Sim* between two users as follows.

$$Sim(u_1, u_2) = \begin{cases} 1, & \text{if } \forall QI \in Q, u_1 == u_2 \\ 0, & \text{if } \forall QI \in Q, u_1 \neq u_2 \\ 0 < Sim(u_1, u_2) < 1, & \text{otherwise} \end{cases} \tag{11}$$

With the help of similarity values, we can obtain a matrix X of highly similar users. Throughout the similarity computation process, user placement can change based on their similarity with other users. In Equation (12), we present an example with 5 users to illustrate this concept.

$$D = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}, X = \begin{pmatrix} u_4 \\ u_3 \\ u_2 \\ u_5 \\ u_1 \end{pmatrix} \tag{12}$$

where D denotes the user arrangement prior to similarity computation, and X denotes the user arrangement/ranking based on the *Sim* values. Afterward, further processing is carried out using only data from X, and users are divided into their respective ECs for further operations.

4.4. Making Equivalence Classes from User Matrix X Based on k

In the literature, there are two types of privacy mechanisms: syntactic and semantic [70]. In the former, users are divided into classes, and anonymization is performed. In the latter type, randomness is introduced into the query answers, and relevant instances are returned with noisy values. Both models have distinct utility and privacy requirements. The work presented here is a syntactic privacy model, and we advance the *k*-anonymity model coupled with uncertainty to anonymize X. The proposed approach divides entire records into multiple classes depending upon the k value and D size. For instance, if $|D| = 1600$ and $k = 50$, then the total number of classes will be 32, whereas each equivalence class encompasses at least k records (e.g., 50 in this case) in it. We find the total number of ECs by using Equation (13).

$$\tau = \frac{|X|(No.\,of\ records)}{k(parameter\,of\ the\ privacy)} \tag{13}$$

The value for k is selected by the data custodian, and it has a significant effect on both utility and privacy. A smaller value for k yields higher utility and lower privacy; a larger value yields the reverse. The selection of the k value is subject to data size, privacy, and utility requirements, the number of data participants, and the mechanism for PPDP. We determine whether the number of ECs will be a whole number (case *i*) or a fraction (case *ii*) using Equation (14).

$$C_u = |X| \ mod\ k \tag{14}$$

where the C_u value can be zero or any other whole number. We use these results to separate both cases above by using Equation (15).

$$Case(i||ii) = \begin{cases} i, & \text{if } C_u == 0 \\ ii, & \text{otherwise} \end{cases} \tag{15}$$

The pseudo-code of the complete algorithm to make ECs based on k from similarity-wise ranked user matrix X, where $X = \{[QI_1, QI_2, \cdots, QI_n], [S]\}$, is given in Algorithm 1.

In Algorithm 1, user matrix X and k are supplied as input, and the set R of the ECs is returned as output. At the start, the number of records is compared with k, and if X contains fewer records than k, the processing stops (line 2). If the number of records is

more than k, the processing continues to divide users into τ classes. Initially, the set R is an empty set. Later, we find the number of ECs using the records in X and the value of k, and the two cases are separated from each other (line 6). Case I is relatively easy because the number of classes is a whole number, and X's partition is made with n ECs, where each EC contains at least k users (lines 7–10). Case II is relatively complex because some records are left over from the assignment process. In this case, we partition X into n ECs, where each EC contains at least k users similar to Case I (lines 12–14). Later, we find the remaining records and assign them to classes by performing a similarity-based proximity analysis (lines 15–16). Afterward, we upgrade the ECs with the new records and add them to set R (lines 17–18). Finally, set R with multiple ECs is returned as the output from Algorithm 1. Each EC can contain at least k users, or $k+i$ users, where $i = 1, 2, 3, 4, 5, \cdots, n$. These classes contain highly similar users with distinct (or the same) values for the SA in each EC. From a privacy point of view, SA values should be fairly distributed in each EC. Moreover, we compute the SA value uncertainty in each EC to find a suitable level for data generalization.

Algorithm 1 Formation of ECs using X and k.

Require: (1) User Matrix X, where $|X| = N$, k.
Ensure: Classes set R while each EC contains at least k users.
1: **if** $(N \leq k)$ **then**
2: Return '*Processing stopped*: less records than k'
3: **else if** $(N > k)$ **then**
4: $R \leftarrow \emptyset$
5: *Compute*, the value of τ and C_u using Equations (13)–(15)
6: **if** $(C_u == 0)$ **then** ▷ # of classes is a whole number
7: **for** $i = 1 : k$: N, where $N = |X|$ do
8: $C_1 = \{X[i], \cdots, X[ik]\}, C_2 = \{X[ik+1], \cdots, X[ik]\}, \ldots, C_n = \{X[(i-1)k], \cdots, X[ik]\}$
9: End for
10: $R \leftarrow \cup\{C_1, C_2, C_3, \ldots, C_n\}$
11: **else if** $C_u \neq 0$) **then** ▷ # of classes is not a whole number
12: **for** $i = 1 : k$: N, where $N = |X|$ do
13: Create ECs with the formula to the possible range: $C_1 = \{X[i], \cdots, X[ik]\}, C_2 = \{X[ik+1], \cdots, X[ik]\}, \ldots, C_n = \{X[(i-1)k], \cdots, X[ik]\}$
14: **End for**
15: **Find** residual records (r_{left})
16: **Assign** the r_{left} to ECs in a way that Sim increases
17: **Upgrade** classes with new records and denote with $C_1^{'}$
18: $R = R \cup \{C_1^{'}, C_2, C_3, \ldots, C_n^{'}\}$
19: **return** R

4.5. Computing Uncertainty in Sensitive Attribute Values in ECs

Computing the uncertainty, U, for each EC is extremely important to effectively address the PUT. To find U, entropy is employed. Entropy is an information theory concept with a wide range of applications in various fields (e.g., physics, thermodynamics, mechanical engineering). In this paper, we use it for measuring the U regarding SA disclosures from ECs. With the help of the U value, we can separate classes where uncertainty is low or high, meaning excessive generalization of the QIs is either needed or not needed at all, respectively. We measure the U value of the SA in C via Equation (16).

$$U(C) = -\sum_{h=1}^{S} p_h log_2 p_h \tag{16}$$

where p_h is the proportion of the SA values in a distinct category, which is computed using (17).

$$p_h = \frac{f_h}{k} \tag{17}$$

where f_h is the frequency of a particular SA value, and k denotes the total number of records in C. With the help of both these equations, U values are computed. The pseudo-code for computing U is given in Algorithm 2.

In Algorithm 2, ECs are provided as the input, and set U of uncertainty values is obtained as the output. To compute the U value from any EC, we use five steps: (i) acquire the class for which the U value needs to be computed, (ii) find all distinct SA values in that EC, (iii) find the frequency of each distinct value, (iv) find the proportion of each distinct value, and (v) calculate the U value using Equation (16). All of these steps are sequentially described in lines 3–7. Finally, we gather and store the U values that will be used in subsequent steps (lines 9–11).

Algorithm 2 Computing the U values of the SA in ECs.

Require: Set R of the equivalence classes.
Ensure: Set U of ECs' uncertainty values.
1: $U \leftarrow \emptyset$
2: **for** $i = 1$ to C_n do
3: Acquire EC of C_i where $U(C_i)$ needs to be computed.
4: Find all distinct values of the SA in a respective C_i.
5: Compute frequency f_h, where $h = \{S_1, S_2, S_3, \ldots, S_n\}$ of each distinct value of the SA.
6: Compute the proportion p of each distinct value of the SA using Equation (17)
7: Determine the uncertainty U of the EC $U(C_1)$ using Equation (16)
8: **End for**
9: Gather computed U values, $\{U(C_1), U(C_2), \ldots, U(C_n)\}$ of each class.
10: Store computed U values in set U, where $U = U \cup \{U(C_1), U(C_2), \ldots, U(C_n)\}$
11: **return** U

In addition to the pseudo-code above, Figure 5 shows the U-computing process with a real example of six records drawn from D.

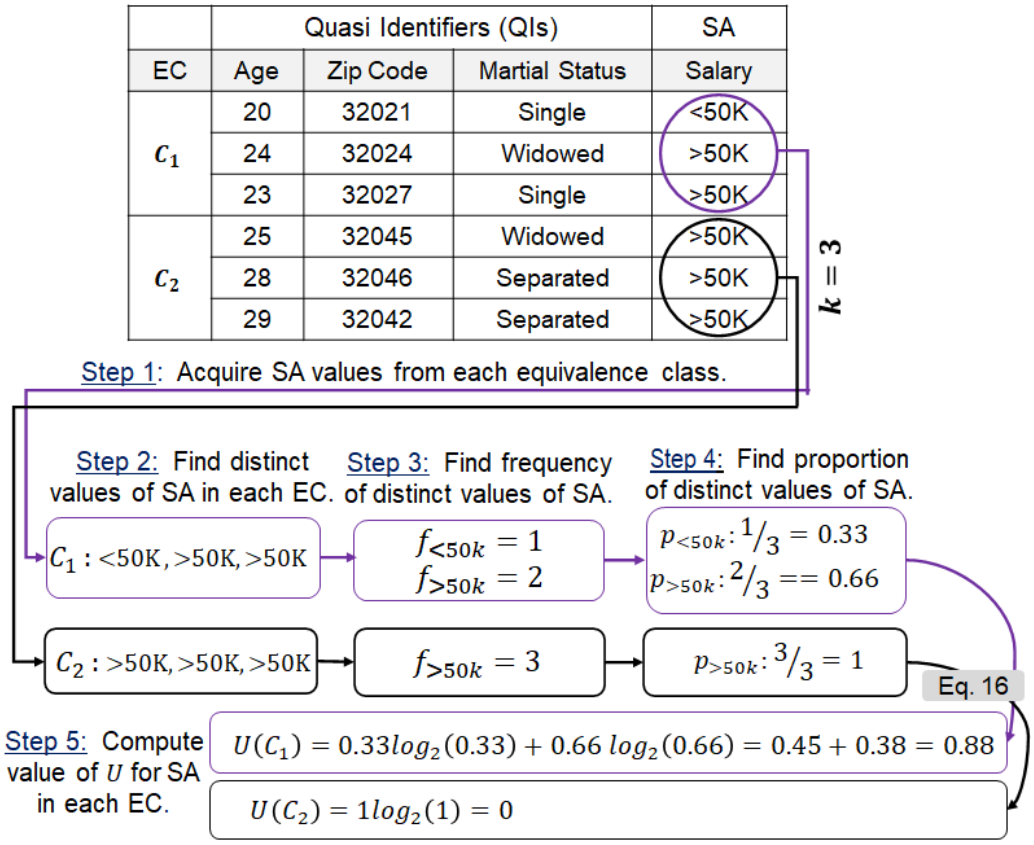

Figure 5. Procedure for computing the uncertainty value U of each EC's SA by leveraging entropy.

In this example, we chose two ECs (with $k = 3$) to demonstrate U's computation process. Since a U value close to 1 is preferred, only C_1 has such a U value. Meanwhile, C_2 has a U value of zero. Therefore, this EC needs ample attention from the SA disclosure point of view. By quantifying the usefulness weights and U values, we can perform only the required generalization, whereas most of the existing methods perform heavy generalization of the data by ignoring these valuable statistics in D.

4.6. Flexible Data Generalization Considering the Usefulness Weight of QI and SA Uncertainty

The final step of the proposed approach is data generalization. In this step, the real QI values are changed to new values that are less precise but consistent in terms of semantics. Generalization is performed with the help of pre-built generalization hierarchies for each QI. We present examples of both numerical and categorical QI generalization hierarchies in Figure 6.

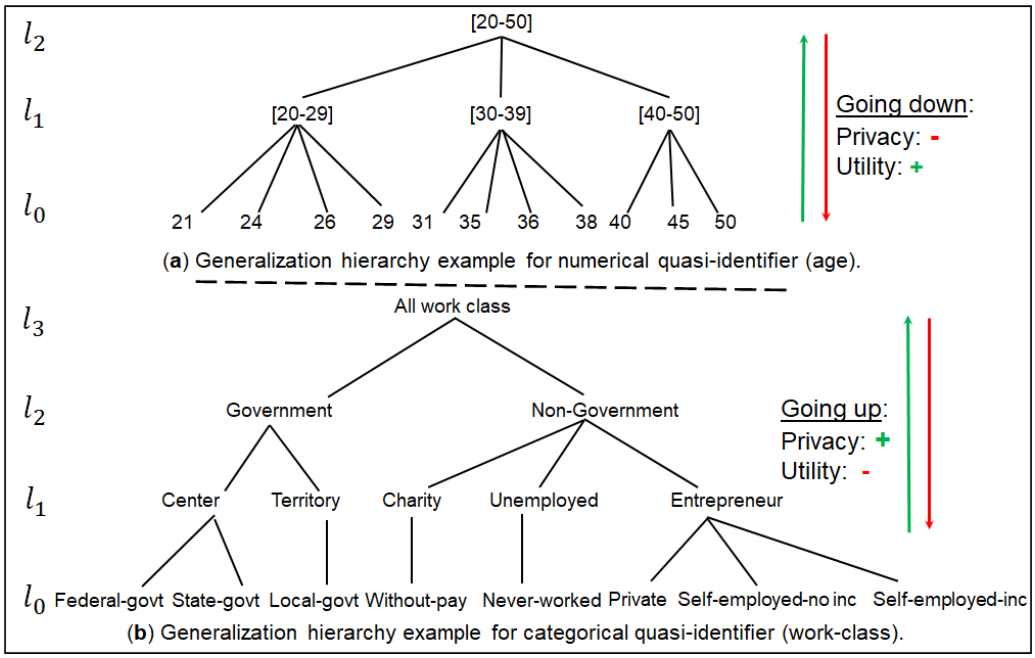

Figure 6. Pictorial overview of the QIs' generalization hierarchies.

There can be n levels in each hierarchy, H_i. Level l_0 denotes the original values of the QIs, and l_n denotes the highest generalization level (also known as suppression). We store all QI hierarchies in set H, where $H = H_{QI_1}, H_{QI_2}, H_{QI_3}, \cdots, H_{QI_n}$, for further processing. Selection of the appropriate generalization level is extremely complex because there exists a strong privacy–utility trade-off, as shown in Figure 6. In lower levels of H, there is maximum utility while privacy is zero, and vice versa. To solve this complex problem, we flexibly perform data generalization by utilizing the intrinsic characteristics of both QIs and the SA. We find the appropriate generalization level from each QI hierarchy that effectively resolves the privacy–utility trade-off in the PPDP. The generalization levels of each hierarchy can be classified into three categories: higher, lower, and intermediate. Since higher levels of any H yield more utility loss, higher-level generalization is only preferred when the SA value has no heterogeneity in an EC (e.g., $U = 0$). Otherwise, generalization is mostly performed on lower or intermediate levels to yield higher accuracy.

By employing similarity-wise EC formation, the U concept, finding and ranking useful QIs, and exploiting other useful intrinsic properties of D, our approach allows for flexible QI generalization, while most existing methods perform generalization based on fixed criteria, resulting in significant utility loss. Furthermore, prior methods sometimes fully or partially hide one or more QIs, significantly limiting published data reusability. In contrast, our proposed approach controls unneeded QI generalization by simultaneously

considering attribute usefulness weight and uncertainty, thereby enhancing data usefulness for practical purposes. In addition, it performs generalization instead of suppression to yield better quality in the anonymized data. In Figure 7, we present the algorithm flowchart that was used to generate $D^{'}$ from D. We segment the whole process into four stages, (i) required data input, (ii) analyzing and comparing U values, (iii) performing the required level generalization, and (iv) combining both categories' results to yield $D^{'}$. In addition to the visual flowchart in Figure 7, we present the pseudo-code of the flexible data generalization procedure in Algorithm 3.

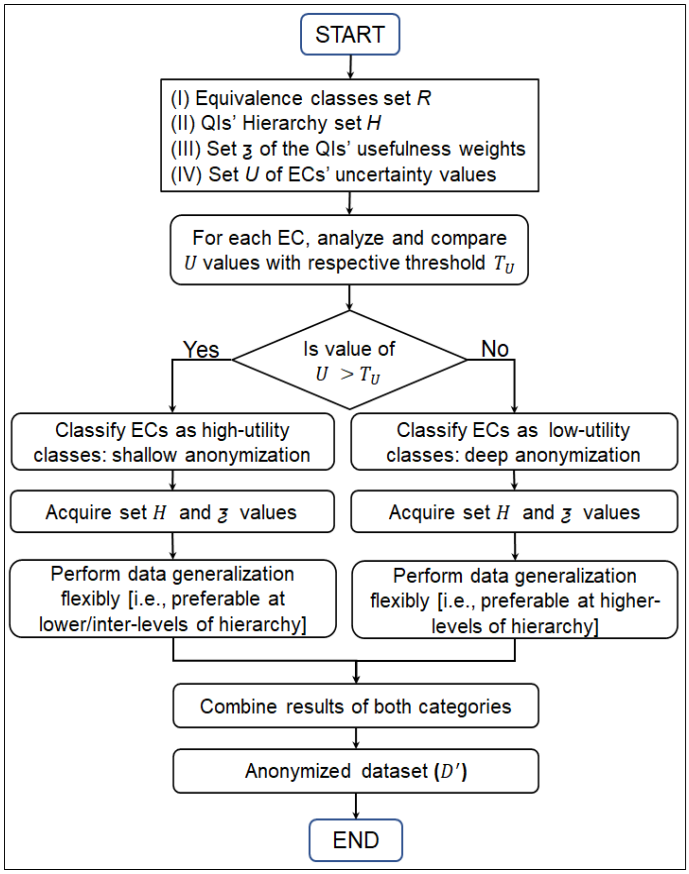

Figure 7. Flowchart of QI usefulness and SA uncertainty-aware flexible data generalization.

In Algorithm 3, the input includes set R of the ECs, the generalization hierarchies set H of the QIs, and QI usefulness weight set ζ. The output is the anonymized dataset $D^{'}$. For each EC, the U values are computed and compared with the corresponding threshold T_U (lines 1–4). The threshold value is chosen through extensive simulations and can be adjusted based on the objectives of data publishing and the PPDP mechanism. Higher and lower utility classes are separated based on the U values. Lines 5–11 perform lower/intermediate-level generalization for ECs with a relatively higher U. In such classes, the majority of the generalization is performed at lower levels of the hierarchies (i.e., the privacy disclosure risk is less). Lines 14–20 perform higher-level generalization for ECs with relatively lower, or zero, U. In such classes, the majority of the generalization is preferred at higher levels of the hierarchies (e.g., the privacy disclosure risk is higher because the SA values lack heterogeneity). Afterward, both categories' results are gathered to produce $D^{'}$ (line 23). Finally, $D^{'}$ is obtained as the output from the algorithm (line 24). $D^{'}$ can be shared with data miners or researchers for secondary uses after performing privacy/utility tests. By flexibly performing data generalization, the semantics of $D^{'}$ are retained as close to the original as possible for precise analysis, except for some ECs.

Algorithm 3 Flexible generalization to produce $D^{'}$.

Require: R, H, ζ
Ensure: $D^{'}$
1: **for** each equivalence class $C_i \in R$ do
2: Compute $U(C_i)$ value of the C_i using Equations (16) and (17).
3: Compare the obtained $U(C_i)$ of the equivalence class C_i with the corresponding T_U.
4: **if** $(U(C_i) > T_U)$ **then**
5: Scenario (*i*): C_i is a high-utility equivalence class (i.e., deep anonymization is not required.)
6: **for** each QI in set Q of an equivalence class C_i
7: Acquire generalization hierarchy H_{QI_1} of QI_1.
8: Obtain usefulness weights Ω_{QI_1} of QI_1 from ζ.
9: Perform QI generalization in a flexible manner ($f^{'}$), preferable at lower or intermediate levels (e.g., l_0, l_1, l_2) of the H_{QI_1}.
10: **Repeat:** $\forall$ QIs, $QI_2, QI_3, \ldots, QI_n$
11: return $f^{'}$
12: **End for**
13: **else if** $(U(C_i) \leq T_U)$ **then**
14: Scenario (*ii*): C_i is a low-utility equivalence class (i.e., deep anonymization is required.)
15: **for** each QI in set Q of an equivalence class C_i
16: Acquire generalization hierarchy H_{QI_1} of QI_1.
17: Obtain usefulness weights Ω_{QI_1} of QI_1 from ζ.
18: Perform QI generalization in a flexible manner ($f^{''}$), preferable at higher levels (e.g., l_{n-2}, l_{n-1}, l_n) of the H_{QI_1}.
19: **Repeat:** $\forall$ QIs, $QI_2, QI_3, \ldots, QI_n$
20: return $f^{''}$
21: **End for**
22: **End for**
23: $D^{'}$ = combine $(f_1^{''} + f_2^{''}+, \ldots, f_n^{''}$ and $f_1^{'} + f_2^{'}+, \ldots, f_n^{'})$.
24: return $D^{'}$

5. Experimental Evaluation

We performed rigorous experiments by utilizing two real-life datasets to benchmark the proposed approach and verify its suitability for practical applications. We compared the performance of our approach with three SOTA anonymization algorithms and models. We present a description of datasets used in the experiments, the hardware/software used the metrics for evaluating the approach's effectiveness, and performance comparisons against the prior solutions in Sections 5.1–5.4.

5.1. Descriptions of Datasets

In the experiments, we considered a relational D encompassing the individuals' identities and SAs. We used the 'Adults' [71] and 'Bkseq' [72] datasets while evaluating and comparing our proposed approach. The former dataset contains four QIs and one SA. Moreover, we ignored non-QI attributes from it. The Bkseq dataset contains three QIs and one SA. Their sizes on the disk are 5.4 MB and 2.85 MB, respectively. We present an overview of the important details of the datasets used in the experimental evaluation in Table 3. Both datasets are openly available [71,72] and have been broadly used for evaluating anonymity solutions.

5.2. Descriptions of the Experimental Environments

All results were performed and compared on a notebook using Windows 10 with a CPU of 2.6 GHz and 8 GB RAM. The results were obtained by utilizing two recognized software packages: Matlab (ver. 9.10.0.1649659 (R2021a)) and RTool for R (ver. 3.6.1 X64 version). Descriptions of the parameters and other useful variables utilized for the QI usefulness weight computations are in Table 4.

Table 3. Description of datasets used for the simulations and comparisons.

Dataset	N	QIs (Distinct Values, # of Levels, Type)	SA (Distinct Values)
Adults [71]	32,561	Age (74, 7, Numerical) Race (5, 3, Categorical) Gender (2, 2, Categorical) Country (41, 4, Categorical)	Salary (2)
Bkseq [72]	16,160	Age (30, 5, Numerical) Weight (30, 3, Numerical) Gender (2, 2, Categorical)	Results of the medical exam (19)

Table 4. Parameters/variables utilized in usefulness weight computations.

Datasets	Parameter Name	Parameter's Values	
		Numerical	Non-Numerical
Adults [71]	Training data size	21,706	-
	Testing data size	10,855	-
	No. of trees (*ntree*)	497	-
	RF model type	-	Classification
	Variable importance	-	true
	Value of *mtry*	4	-
	Predictors	-	All QIs
	Target class	-	Salary
	Keep forest	-	true
Bkseq [72]	Training data size	10,773	-
	Testing data size	5,387	-
	No. of trees (*ntree*)	267	-
	RF model type	-	Classification
	Variable importance	-	true
	Value of *mtry*	3	-
	Predictors	-	All QIs
	Target class	-	Medical exam result
	Keep forest	-	true

The parameter values were determined from rigorous experiments and analysis. The training data constituted $^2/_3$ of D, and testing data were $^1/_3$ of D. In addition, the target variable was categorical; this is why the RF model type is classification rather than regression. Table 5 presents the ζ of QIs present in the D presented in Table 3. The symbol '-' shows that the particular QI does not belong. These weights were calculated by RF through repeated tests. We determined and used the optimal values of RF parameters (as listed in Table 4) while computing the QI usefulness weights.

Table 5. Usefulness weight values of the QIs in both datasets.

Dataset	Quasi-Identifiers and Their Usefulness Weights				
	Age	Gender	Race	Country	Weight
Adults [71]	0.18	1.10	26.91	71.81	-
Bkseq [72]	47.51	38.91	-	-	13.58

The highest usefulness weight being assigned to the QI country is justified by the fact that in the Adult dataset, the majority of records have the U.S. value in the country column. The occurrence of a country value other than the U.S. is rare and only happens when the k value is very small. As a result, the country QI has a high usefulness weight and a lower impact on individual privacy. QIs that are not concentrated on a specific value, such as age, have relatively lower usefulness weights. We validated these weights by analyzing the distribution of each QI's original values in D. Similarly, in the Bkseq dataset, the age QI has less variability, and therefore its weight value is higher compared to the other QIs. In the Bkseq dataset, the usefulness weight value tendency is not very high for one QI due to the higher distinct values of the SA (e.g., 19) compared to the Adult dataset (e.g., 2). We performed validation to ensure the correctness of the weight values determined by RF by

assessing the original values and their domains. From the validation results, we found that values computed by RF are highly accurate and reliable for practical tasks.

5.3. Descriptions of Metrics and Evaluation Criteria

To evaluate and compare the efficacy of our approach, we utilized three metrics. Two metrics were used to measure the utility of the $D^{'}$, and one metric was used for privacy evaluation. To evaluate anonymous data utility, we used information loss (IL) and classification/regression model accuracy. IL belongs to the general-purpose metrics category for utility estimation. To calculate IL, we employed distortion measure (DM), which is stable, and the most widely used IL metric. Fung et al. [73] explained the DM metric in their study. DM values can be computed by analyzing the hierarchy levels upon which QI values are transformed. The value of the DM metric is computed with Equation (18).

$$DM = \sum_{v=n_1}^{u_N} \sum_{q=1}^{n} \frac{l_i}{l_t} \times \zeta_q \tag{18}$$

where l_i denotes the level of the hierarchy on which the QI value is transformed, and l_t denotes the total number of levels in H, while ζ_q denotes the usefulness weight of the QI. If a QI value is not generalized, then the value of DM will be zero.

Accuracy belongs to the category of special-purpose metrics and has been extensively used to evaluate the $D^{'}$ quality for mining/analytical purposes. Generally, a higher value of accuracy is desirable for informative analytics of published data. To achieve superior accuracy, the domain consistency in the QI values is important when generalizing the data. The value of accuracy can be determined using the ML methods (i.e., decision trees, RF, support vector machines, etc.) via Equation (19).

$$Accuracy = \frac{True\ negatives + True\ positives}{Total\ number\ of\ users\ in\ D'} \tag{19}$$

To evaluate privacy protection, we used a privacy-sensitive pattern (PSP)-based probabilistic disclosure (PD) metric. The PSP antecedent is the QI's dominant value, and the consequent is the SA value. In general, it captures the association between the SA and the QIs as r_i,$((QI_{1_v}, QI_{2_v}, QI_{3_v}) \rightarrow s_i)$. Analysts (or attackers) can make many of these PSPs by analyzing the QI domain values or from ancillary information obtained from exterior sources to compromise a user's identity/SA. Very high PD values indicate that many patterns can be derived with ease by data mining firms or attackers, and accordingly, privacy protection is insufficient. We made multiple PSPs and quantified the PD to perform a quantitative analysis of privacy protection. The PD of the *ith* PSP in the *jth* EC can be computed using Equation (20).

$$P(C_j(r_i)) = \frac{\sigma}{k} \tag{20}$$

where σ denotes the correct records that match the antecedents and the consequences of a PSP and k denotes the total number of users in the respective EC.

All three metrics (DM, accuracy, and PD) were used to measure the effectiveness of the proposed approach. For fair comparisons, we prepared anonymized versions of both datasets with varying scales of k. We used both large-scale values (L_s, where $L_s = D^{'}_k$ and $k = [50, 75, 100, 125, 150, 175, 200, 250]$), and small-scale values (S_s, where $S_s = D^{'}_k$ and $k = [5, 10, 15, 20, 25, 30, 35, 40]$) of k to demonstrate the potency of the proposed approach.

5.4. Performance against Existing Anonymization Algorithms and Models

To evaluate the proposed approach, we compared the performance with three prior anonymizing methods: IACk [39], WFDA [52], and CPA [57]. All three anonymization algorithms are competitive in addressing competing goals of utility and privacy. Furthermore, to demonstrate the efficacy of our anonymizing approach for practical applications, we used three privacy models as a baseline in our simulation experiments and compared

results with them. All three privacy models, i.e., k-anonymity [29], ℓ-diversity [30], and t-closeness [31], have higher adoption rates for person-specific data anonymization. These models perform anonymization mainly by enforcing constraints on either QI or SA values. By not utilizing the usefulness weights concept and uncertainty combined with flexible generalization, their flaws have been recognized in a large number of studies. In the following subsections, we present quantitative results obtained through a series of experiments.

5.4.1. Comparisons of Anonymous Data Utility/Quality

In this subsection, we evaluate and compare the performance of our approach based on two metrics: special purpose (i.e., accuracy) and general purpose (i.e., IL). Formalizations for both metrics are given in Equations (19) and (18), respectively.

(i) Comparisons of Accuracy: The first analysis involved accuracy computation and a comparison with prior solutions. We experimented on two datasets using S_s and L_s values of k. From each version of the anonymized datasets, accuracy values were obtained and compared with the existing solutions. All results were obtained using the R programming language. We present the accuracy comparison with varying k scales in Figure 8. With increases in k values, accuracy values also increased. Moreover, the proposed approach performed consistently better compared to the previous methods for most k values. The accuracy values of our anonymization approach were only marginally lower than D.

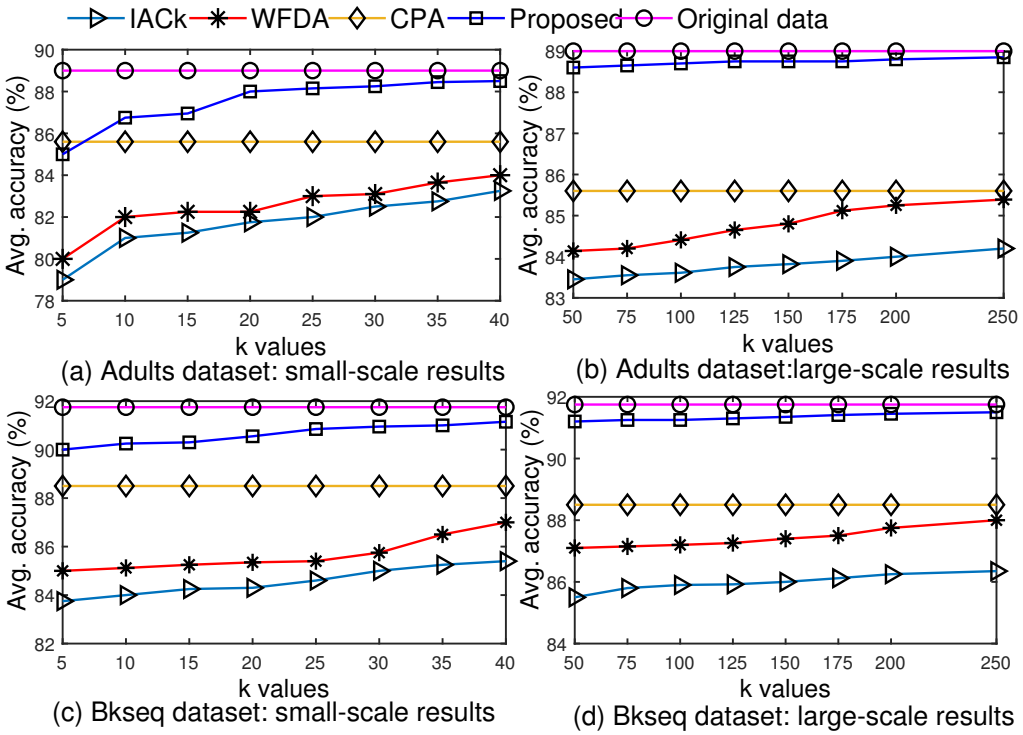

Figure 8. Accuracy: proposed approach versus existing algorithms and original data.

The proposed approach, on average, enhanced accuracy by 9.81% with the Adult dataset, and by 10.66% with the Bkseq dataset. These results emphasize the proposed approach's feasibility in terms of better data quality for data mining and analytical purposes. To further validate the suitability of the proposed approach, simulation results were also compared with three privacy models. The average accuracy results obtained from the experiments and comparisons with existing anonymization models are shown in Table 6.

Table 6. Average accuracy: the proposed approach versus existing models.

Dataset	Anonymization Mechanisms (e.g., the Proposed Approach and Existing Privacy Models)				
	***k*-Anonymity [29]**	**ℓ-Diversity [30]**	***t*-Closeness [31]**	**Our Approach**	**Original Data**
Adults [71]	86.46	85.71	84.09	88.71	89.01
Bkseq [72]	88.16	87.91	86.26	91.64	91.75

The proposed approach, on average, yielded better results, compared to the existing models, with both datasets. The proposed approach had an 8.61% improvement in accuracy compared to the existing models. To further enhance the persuasiveness of our work, we compared the results with the recent SOTA ML-based anonymization method. The results obtained from the two benchmark datasets and their comparisons are given in Table 7. From the results, it can be seen that our approach yielded higher accuracy compared to the SVD3RD method. These results fortify the significance of our approach for data mining tasks.

Table 7. Average accuracy: the proposed approach versus the existing SOTA method.

Dataset	Anonymization Mechanism and Original Data		
	SVD3RD [59]	**Our Approach**	**Original Data**
Adults [71]	77.10	88.71	89.01
Bkseq [72]	83.89	91.64	91.75

(ii) Comparisons of Information Loss: IL is an unavoidable and unfortunate aftereffect of any anonymizing operation applied to D.

To yield lower IL values, anonymizing needs to be performed carefully by controlling higher-level generalization (i.e., levels l_{n-2}, l_{n-1}, l_n) to the greatest extent possible. In some cases, $D^{'}$ retains no usefulness, and extensive post-processing is needed for any analysis. In addition, over-anonymized data become useless to data miners. To resolve such issues of unnecessary transformations, our approach performs minimal generalization on the data. Higher-level generalization is avoided and performed only when there is a greater chance of SA disclosure. By performing generalization in a flexible manner and restricting it to the lowest possible level of H, IL can be significantly reduced. Figure 9 presents the IL values obtained from experiments conducted with different k scales.

From the results, we can see that IL values consistently increase with increases in k. Moreover, the IL values produced by our approach are less than the existing methods. The proposed anonymization approach, on average, reduced IL by 11.36%, compared to sophisticated, closely related algorithms. These results obtained are due to flexible generalization, whereas the existing algorithms fixedly perform generalization (e.g., one generalization level in all ECs). These simulation results emphasize the validity of the proposed approach for general-purpose applications. To further assess the capabilities of the proposed anonymization approach, we compared the results with three privacy models with 10 representative k values (e.g., $k = \{5, 10, 20, 30, 40, 50, 100, 150, 200, 250\}$). The results and a comparison with existing state-of-the-art privacy models are shown in Figure 10.

From the results, we can see that IL increased with an increase in k-values for both datasets. Moreover, the proposed approach had lower IL values in most cases, except the first two. In those cases (i.e., when k was small), most ECs had a lower U in the SA values. Hence, higher-level anonymization was performed to curtail privacy breaches. However, from an IL point of view, the proposed approach, on average, showed a 29.6% reduction in IL compared to the prior privacy models. In most tests, a lower IL is due to flexible generalization, and from using the similarity concept while making the ECs.

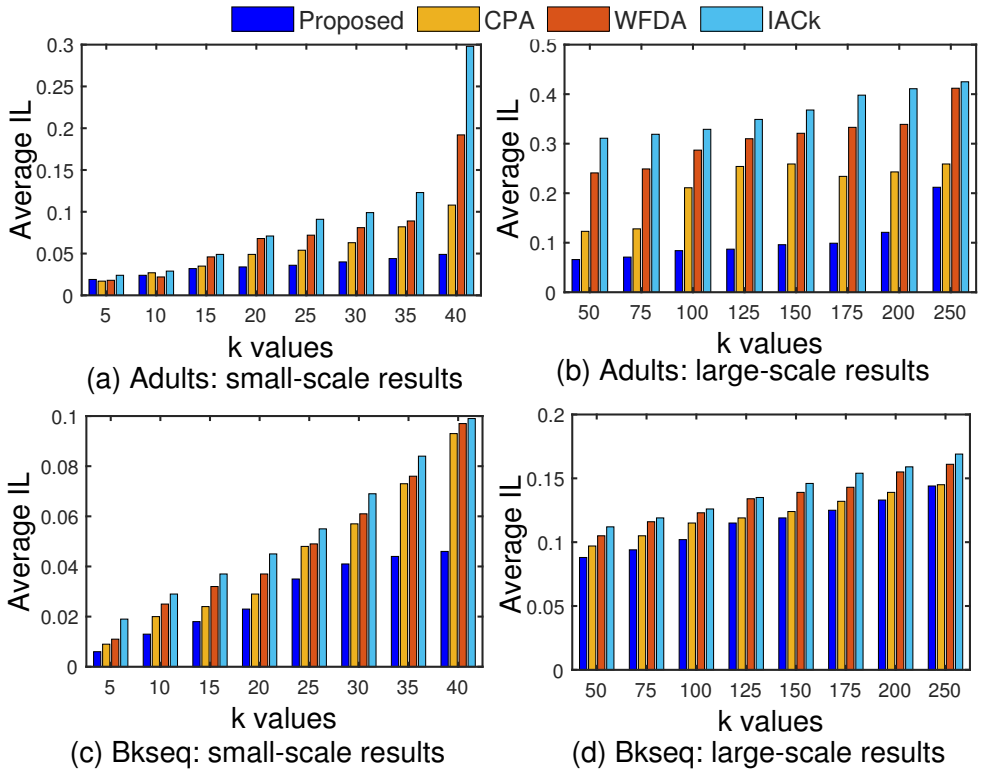

Figure 9. Information loss: Proposed approach versus existing algorithms.

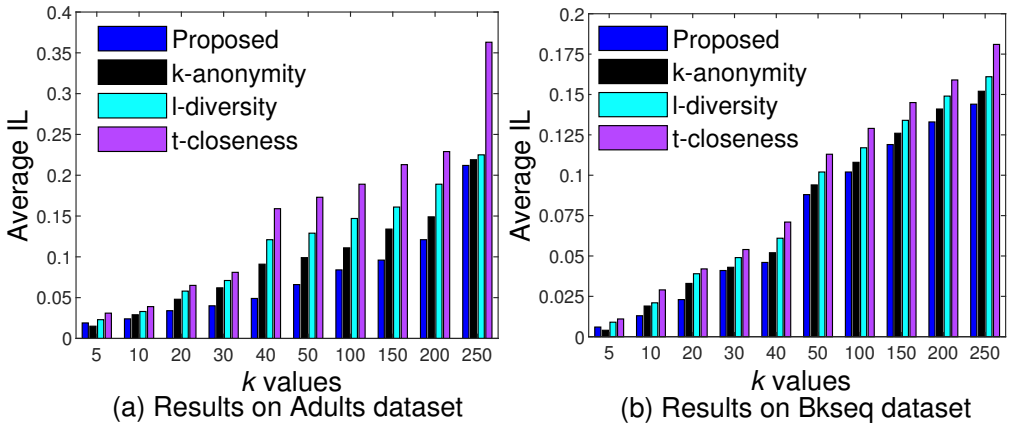

Figure 10. Information loss: Proposed approach versus existing models.

5.4.2. Comparison of Individual Privacy Preservation

Comparisons of Probabilistic Disclosure: Although the main assertion of the proposed approach is utility enhancement, privacy evaluation is still imperative because of the trade-off between privacy and utility. We evaluated the essence of the proposed approach from the perspective of privacy preservation and compared the results with existing models and algorithms. We made seventeen different versions of each dataset and performed PSP-based PD analysis. We created different anonymized versions of each dataset based on the k value. The anonymized version of the same dataset created with $k = 5$ is different from $k = 2$. For example, if $k = 5$, the number of users in each class is at least 5, and the total number of classes can be determined by dividing $|D|$ by 5. In contrast, if $k = 2$, the number of users in each class is at least 2, and the total number of classes can be determined by dividing $|D|$ by 2. Each version of the anonymized dataset can be differentiated from others based on the number of classes as well as the number of users in each class. Specifically, k can help in differentiating the anonymized version of datasets from each other. Furthermore,

we performed a worst-case analysis by choosing more dominant values of the distinct QIs and determined the correct matches. The value of PD ranges between 0 and 1. A value of 1 means there is 100% disclosure, based on the QI values; a value of zero indicates no disclosure. We constructed multiple patterns and evaluated effectiveness in terms of possible correct matches. The results are shown in Table 8.

Table 8. Average PD values: proposed approach versus existing solutions.

Dataset	k	PD Comparisons with Existing Studies with Varying k Values		
		Existing Algorithms	**Existing Models**	**Proposed Approach**
Adults [71]	2	0.87	0.67	0.54
	5	0.60	0.68	0.42
	10	0.63	0.69	0.51
	15	0.75	0.74	0.64
	20	0.69	0.75	0.66
	25	0.80	0.78	0.72
	30	0.78	0.79	0.67
	35	0.80	0.84	0.73
	40	0.81	0.84	0.75
	50	0.83	0.85	0.76
	75	0.84	0.83	0.78
	100	0.82	0.85	0.77
	125	0.86	0.86	0.79
	150	0.88	0.89	0.80
	175	0.88	0.91	0.81
	200	0.89	0.92	0.82
	250	0.88	0.94	0.82
Bkseq [72]	2	0.50	0.50	0.50
	5	0.34	0.38	0.26
	10	0.39	0.41	0.33
	15	0.53	0.55	0.42
	20	0.57	0.63	0.48
	25	0.65	0.69	0.53
	30	0.63	0.71	0.54
	35	0.67	0.73	0.63
	40	0.69	0.74	0.64
	50	0.71	0.76	0.65
	75	0.73	0.79	0.67
	100	0.75	0.81	0.69
	125	0.76	0.82	0.71
	150	0.77	0.81	0.71
	175	0.79	0.84	0.73
	200	0.81	0.85	0.75
	250	0.81	0.86	0.77

The PD values of the proposed approach are lower in most cases compared to existing methods and models. The proposed approach performs poorly when k values are very small. However, in real-world cases, the values of k are kept relatively large. With the Adult dataset, the proposed approach lowered PD values by 12.51%, compared to existing solutions. In contrast, in the Bkseq dataset, our approach reduced PD values by 9.01%. These results demonstrate the suitability of the proposed approach for effective privacy preservation. Apart from the empirical results, the proposed approach assists in anonymizing skewed datasets, which is practically impossible with most of the anonymizing solutions that rely on heterogeneous values in each group/EC. The proposed approach is an offline approach, and it has significantly lower space and time complexities with the pre-computed values of three statistics (e.g., Sim, U, and ζ). In addition, it can be applied to both categorical and numerical datasets. The proposed approach is generic and

can be highly applicable in the medical domain where information seekers usually expect high-quality $D^{'}$ for research purposes.

Comparisons of re-identification risk in homogeneous attacks: In the syntactic methods (e.g., methods that create classes and apply either generalization or suppression to anonymize data), there is a risk that all records in an EC can have the same SA in the case of imbalanced datasets, leading to higher re-identification risks. For example, in the Adult dataset, there are only two values of the SA, and the distribution is highly skewed. To further highlight the trade-off between the privacy/risk and utility of the proposed approach, we employed the re-identification risk against homogeneous attacks [50]. The $\mathcal{HA}$ value can be computed using Equation (21).

$$\mathcal{HA} = \frac{1}{N} \sum_{j \in R} f_j \times I \tag{21}$$

The $\mathcal{HA}$ is the ratio of records with identical SA values in an EC to all records in the D. We computed the $\mathcal{HA}$ from the different versions of the anonymized datasets, and the results are given in Table 9. Specifically, we present the total number of classes with identical SA, and the average $\mathcal{HA}$ value for each dataset.

Table 9. Average $\mathcal{HA}$ values: proposed approach versus existing solutions.

Dataset	k (# of Classes)	$\mathcal{HA}$'s Comparisons with Existing Studies with Varying k Values		
		Existing Algorithms	Existing Models	Proposed Approach
Adults [71]	2 (16,280)	4549.56	6210.5	2779.5
	5 (6512)	1819.90	2484.2	1111.8
	10 (3256)	909.81	1242.10	555.81
	15 (2171)	606.53	828.06	370.61
	20 (1628)	454.32	621.05	277.95
	25 (1302)	363.92	496.84	222.36
	30 (1085)	303.26	414.03	185.31
	35 (930)	259.94	354.82	158.82
	40 (814)	227.45	310.52	138.97
	50 (651)	181.96	248.42	111.18
	75 (434)	121.30	165.61	74.12
	100 (326)	90.98	124.21	55.60
	125 (260)	72.78	99.36	44.48
	150 (217)	60.65	82.80	37.06
	175 (186)	51.98	79.79	31.76
	200 (163)	45.49	62.10	27.80
	250 (130)	36.39	49.64	22.36
Average $\mathcal{HA}$ value		0.32	0.42	0.19
Bkseq [72]	2 (8080)	2050.21	2600.21	1450.32
	5 (3232)	820.21	1040.29	580.92
	10 (1616)	410.34	520.31	290.32
	15 (1077)	273.33	346.67	193.21
	20 (808)	205.10	260.89	145.34
	25 (646)	164.32	208.32	116.45
	30 (539)	136.66	173.33	96.67
	35 (462)	117.14	148.57	82.85
	40 (404)	102.51	130.43	72.55
	50 (323)	82.32	104.65	58.98
	75 (215)	54.66	69.43	38.66
	100 (162)	41.34	52.60	29.87
	125 (129)	32.81	41.67	23.24
	150 (108)	27.33	34.66	19.35
	175 (92)	23.42	29.82	16.59
	200 (81)	20.51	26.69	14.43
	250 (65)	16.41	20.89	11.65
Average $\mathcal{HA}$ value		0.14	0.16	0.09

From the results, it can be seen that the number of classes with identical SAs decreases when k increases. The Adult dataset has a higher imbalance compared to Bkseq and, therefore, the number of classes with identical SAs is relatively higher than in the Bkseq dataset. Due to the higher imbalance in the Adult dataset, the average $\mathcal{HA}$ value is also higher. In contrast, the Bkseq dataset is balanced and contains many diverse values for SA; therefore, the average $\mathcal{HA}$ values are small. From the models, k-anonymity has a lower performance than other models. In the algorithms, IACk exhibits a deficient performance by not considering diversity in the SA column. The proposed approach demonstrates lower re-identification risk in homogeneous attacks compared to existing solutions. The main reason for this is the consideration of U in each EC, whereas most existing solutions often ignore uncertainty in the SA column, leading to higher re-identification risk. These results reinforce the significance of our approach in terms of better safeguarding against re-identification risk.

Analysis of the memory/storage trade-offs required to run the proposed algorithms: The proposed approach has acceptable space complexity, even when entire datasets were loaded into memory, and no out-of-memory error occurred during experimentation. All steps of the proposed approach were executed in a pipeline fashion and, thus, memory/storage trade-offs were effectively resolved. Additionally, having access to entire datasets, generalization hierarchies, and usefulness weights in advance minimized memory consumption. However, in some steps (such as similarity and weight computation), memory consumption was relatively higher than in other steps. To reduce resource utilization, the memory consumption in weight computation was reduced by optimizing hyperparameters by leveraging the grid search function. Furthermore, the NSAs were eliminated at the outset to reduce resource utilization. In the similarity computation process, the upper triangular part of the similarity matrix, which is merely a replica of the lower triangular, was also removed to reduce memory consumption. In all other steps, the space complexity was manageable. However, the space complexity of the proposed approach can increase when the size of the dataset increases vertically, horizontally, or both; therefore, ample attention is required to resolve the memory/storage trade-offs. To run the proposed approach on resource-constrained devices and to prevent memory issues, all six steps can be executed sequentially.

On the generalizability and applicability of the proposed approach in big data contexts: A key question that arises about the results obtained through experiments is *how valid and generalizable are the results?* To answer this question concerning the experimental results produced by our approach, we evaluated the threats to validity in terms of internal, external, statistical, and construct validity. Based on detailed analysis, it can be concluded that threats to the validity and generalizability of our method are limited based on the following rationales: (i) most parameter values are adaptive and flexible and, therefore, the obtained results are general and universal; (ii) the D used in the experiments are large enough and contain QIDs of both types (numerical and categorical), the experimental results are reliable, and the conclusions are credible and valid; (iii) through a fair comparison and analysis of our method with prior state-of-the-art methods and models, it significantly outperforms them in enhancing utility and preserving privacy; (iv) the experimental setup is highly similar to real-world environments, which can augment the use of our method in many commercial applications. The proposed approach can be applied to any real-world dataset of higher dimensions. However, the space and time complexity can rise when the proposed approach is applied in Big Data contexts. We analyzed the memory and time performances of our approach by creating small-sized and large-sized chunks of the available data. The experimental analysis shows that both time and space complexity do not grow dramatically and, therefore, our approach is scalable when applied in Big Data scenarios. However, careful attention is needed when NSAs are not removed from data, and the number of QIs is large (in numbers). The complications of our approach can be lowered further by using pre-computed statistics of some steps (e.g., usefulness weights, similarity matrix, the uncertainty of SA in classes, and pre-built generalization hierarchies) in Big Data contexts.

6. Conclusions and Future Work

In this paper, we proposed a novel and generic approach for anonymizing person-specific data. The main objectives of the proposed approach are to enhance the reusability of the anonymized data for analytical purposes (e.g., understanding disease dynamics, trend predictions, cause-and-effect relationship analysis, and demography-based knowledge extraction) while providing considerable privacy. We proposed a mechanism for computing the usefulness weights of each quasi-identifier by using the random forest to limit heavy changes in useful QIs during data anonymization. We employed the concept of information theory for calculating the uncertainty of SA values in classes to minimize the privacy breaches caused by classes of low uncertainty. Furthermore, the proposed flexible generalization method anonymizes person-specific data, considering the inherent statistics (e.g., usefulness weights and uncertainty) of attributes from the original data. It resolves the utility issues that arise from fixed-manner anonymization (i.e., by not identifying information-rich attributes) while sustaining considerable privacy in the anonymized data. We conducted numerous experiments on two real-life and benchmark datasets to assess the suitability of our proposed approach for real-world applications. The experimental results confirm the superiority of our approach over existing SOTA methods and models. The proposed anonymization approach, on average, enhanced the accuracy by 9.81% with the Adult dataset, and by 10.66% with the Bkseq dataset compared to the SOTA anonymization algorithms. The proposed approach had an 8.61% improvement in accuracy, compared to the existing privacy models. Furthermore, the accuracy values from our approach are only marginally lower (i.e., 0.30% (Adult dataset) and 0.11% (Bkseq dataset)) than the original data. From the IL point of view, the proposed approach has shown 11.36% and 29.6% improvements compared to existing algorithms and models, respectively. The proposed approach has significantly reduced the disclosure risk compared to previous solutions. With the Adult dataset, the proposed approach lowered the disclosure values by 12.51%, compared to existing solutions. In contrast, in the Bkseq dataset, our approach reduced the disclosure values by 9.01%. These results fortify the significance of the proposed approach from the perspective of effective privacy preservation without losing guarantees of anonymous data utility. We intend to extend the proposed approach for optimizing the privacy–utility trade-off by incorporating usefulness and uncertainty concepts in the objective function of clustering techniques. We intend to apply the proposed approach on large-size datasets, such as IHIS (https://www.ihis.com.sg/, accessed on 5 April 2023), to verify the scalability in realistic scenarios. In addition, we plan to amalgamate DP with the proposed approach to further improve the privacy and utility results.

Author Contributions: Conceptualization, A.M. and S.O.H.; methodology, A.M.; software, A.M.; validation, A.M. and S.O.H.; formal analysis, A.M.; investigation, A.M. and S.O.H.; resources, A.M. and S.O.H.; data curation, A.M.; writing—original draft preparation, A.M.; writing—review and editing, A.M. and S.O.H.; visualization, A.M.; supervision, S.O.H.; project administration, S.O.H.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Agency for Defense Development by the Korean Government (UI220040XD).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in the experimental evaluation of this study are available within this article.

Acknowledgments: We thank the four expert reviewers who thoroughly evaluated this paper and provided very constructive feedback, which significantly enhanced the technical depth of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Parvinen, L.; Alamäki, A.; Hallikainen, H.; Mäki, M. Exploring the challenges of and solutions to sharing personal genomic data for use in healthcare. *Health Inform. J.* **2023**, *29*, 14604582231152185. [CrossRef] [PubMed]
2. O'Leary, J.C. Data sharing: The public's perspective. In *Genomic Data Sharing*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 157–170.
3. Li, F.; Li, H.; Niu, B.; Chen, J. Privacy computing: Concept, computing framework, and future development trends. *Engineering* **2019**, *5*, 1179–1192. [CrossRef]
4. Tran, H.Y.; Hu, J. Privacy-preserving big data analytics a comprehensive survey. *J. Parallel Distrib. Comput.* **2019**, *134*, 207–218. [CrossRef]
5. Majeed, A.; Hwang, S.O. Quantifying the Vulnerability of Attributes for Effective Privacy Preservation Using Machine Learning. *IEEE Access* **2023**, *11*, 4400–4411. [CrossRef]
6. Jayabalan, M.; Rana, M.E. Anonymizing healthcare records: A study of privacy preserving data publishing techniques. *Adv. Sci. Lett.* **2018**, *24*, 1694–1697. [CrossRef]
7. Akinkunmi, O.; Rana, M.E. Privacy preserving data publishing anonymization methods for limiting malicious attacks in healthcare records. *J. Comput. Theor. Nanosci.* **2019**, *16*, 3538–3543.
8. Su, B.; Huang, J.; Miao, K.; Wang, Z.; Zhang, X.; Chen, Y. K-Anonymity Privacy Protection Algorithm for Multi-Dimensional Data against Skewness and Similarity Attacks. *Sensors* **2023**, *23*, 1554. [CrossRef]
9. Yağar, F. Growing Concern During the COVID-19 Pandemic: Data Privacy. *Turk. Klin. J. Health Sci.* **2021**, *6*, 387–392. [CrossRef]
10. Jian, X.; Wang, W.; Pei, J.; Wang, X.; Shi, B.; Fu, A.W.C. Utility-based anonymization using local recoding. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 20–23 August 2006; pp. 785–790.
11. Xu, J.; Wang, W.; Pei, J.; Wang, X.; Shi, B.; Fu, A.W.C. Utility-based anonymization for privacy preservation with less information loss. *ACM Sigkdd Explor. Newsl.* **2006**, *8*, 21–30. [CrossRef]
12. Onesimu, J.A.; Karthikeyan, J.; Eunice, J.; Pomplun, M.; Dang, H. Privacy Preserving Attribute-Focused Anonymization Scheme for Healthcare Data Publishing. *IEEE Access* **2022**, *10*, 86979–86997. [CrossRef]
13. Lin, C.Y. A reversible privacy-preserving clustering technique based on k-means algorithm. *Appl. Soft Comput.* **2020**, *87*, 105995. [CrossRef]
14. Li, T.; Li, J.; Chen, X.; Liu, Z.; Lou, W.; Hou, T. NPMML: A framework for non-interactive privacy-preserving multi-party machine learning. *IEEE Trans. Dependable Secur. Comput.* **2020**, *18*, 2969–2982. [CrossRef]
15. Wang, R.; Zhu, Y.; Chang, C.C.; Peng, Q. Privacy-preserving high-dimensional data publishing for classification. *Comput. Secur.* **2020**, *93*, 101785. [CrossRef]
16. Eicher, J.; Bild, R.; Spengler, H.; Kuhn, K.A.; Prasser, F. A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models. *BMC Med Inform. Decis. Mak.* **2020**, *20*, 1–14. [CrossRef]
17. Brough, A.R.; Martin, K.D. Consumer privacy during (and after) the COVID-19 pandemic. *J. Public Policy Mark.* **2021**, *40*, 108–110. [CrossRef]
18. Foraker, R.E.; Lai, A.M.; Kannampallil, T.G.; Woeltje, K.F.; Trolard, A.M.; Payne, P.R. Transmission dynamics: Data sharing in the COVID-19 era. *Learn. Health Syst.* **2021**, *5*, e10235. [CrossRef]
19. Lenert, L.; McSwain, B.Y. Balancing health privacy, health information exchange, and research in the context of the COVID-19 pandemic. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 963–966. [CrossRef]
20. Strobel, M.; Shokri, R. Data Privacy and Trustworthy Machine Learning. *IEEE Secur. Priv.* **2022**, *20*, 44–49. [CrossRef]
21. He, Z.; Cai, Z.; Yu, J. Latent-data privacy preserving with customized data utility for social network data. *IEEE Trans. Veh. Technol.* **2017**, *67*, 665–673. [CrossRef]
22. Majeed, A.; Hwang, S.O. Rectification of Syntactic and Semantic Privacy Mechanisms. *IEEE Secur. Priv.* **2022**, *1*, 2–16. [CrossRef]
23. Mohammed, N.; Chen, R.; Fung, B.C.; Yu, P.S. Differentially private data release for data mining. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 493–501.
24. Dwork, C. Differential privacy: A survey of results. In Proceedings of the International Conference on Theory and Applications of Models of Computation, Xi'an, China, 25–29 April 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 1–19.
25. Li, Y.; Liu, Y.; Li, B.; Wang, W.; Liu, N. Towards practical differential privacy in data analysis: Understanding the effect of epsilon on utility in private erm. *Comput. Secur.* **2023**, *128*, 103147. [CrossRef]
26. Li, Y.; Li, B.; Wang, W.; Liu, N. An Efficient Epsilon Selection Method for DP-ERM with Expected Accuracy Constraints. In Proceedings of the 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Wuhan, China, 9–11 December 2022; pp. 533–540.
27. Domingo-Ferrer, J.; Sánchez, D.; Blanco-Justicia, A. The limits of differential privacy (and its misuse in data release and machine learning). *Commun. ACM* **2021**, *64*, 33–35. [CrossRef]
28. Singh, R.; Dwivedi, A.D.; Srivastava, G.; Chatterjee, P.; Lin, J.C.W. A Privacy Preserving Internet of Things Smart Healthcare Financial System. *IEEE Internet Things J.* **2023**. [CrossRef]
29. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570. [CrossRef]

30. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3-es. [CrossRef]
31. Li, N.; Li, T.; Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 17–20 April 2007; pp. 106–115.
32. Sun, X.; Sun, L.; Wang, H. Extended k-anonymity models against sensitive attribute disclosure. *Comput. Commun.* **2011**, *34*, 526–535. [CrossRef]
33. Chen, L.; Zhong, S.; Wang, L.e.; Li, X. A Sensitivity-Adaptive ρ-Uncertainty Model for Set-Valued Data. In Proceedings of the International Conference on Financial Cryptography and Data Security, Christ Church, Barbados, 22–26 February 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 460–473.
34. Wong, R.C.W.; Li, J.; Fu, A.W.C.; Wang, K. (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 20–23 August 2006; pp. 754–759.
35. Sun, X.; Li, M.; Wang, H. A family of enhanced (L, α)-diversity models for privacy preserving data publishing. *Future Gener. Comput. Syst.* **2011**, *27*, 348–356. [CrossRef]
36. Soria-Comas, J.; Domingo-Ferrer, J.; Sanchez, D.; Martinez, S. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 3098–3110. [CrossRef]
37. Ashkouti, F.; Sheikhahmadi, A. DI-Mondrian: Distributed improved Mondrian for satisfaction of the L-diversity privacy model using Apache Spark. *Inf. Sci.* **2021**, *546*, 1–24. [CrossRef]
38. Zigomitros, A.; Casino, F.; Solanas, A.; Patsakis, C. A survey on privacy properties for data publishing of relational data. *IEEE Access* **2020**, *8*, 51071–51099. [CrossRef]
39. Li, J.; Liu, J.; Baig, M.; Wong, R.C.W. Information based data anonymization for classification utility. *Data Knowl. Eng.* **2011**, *70*, 1030–1045. [CrossRef]
40. Cagliero, L.; Garza, P. Improving classification models with taxonomy information. *Data Knowl. Eng.* **2013**, *86*, 85–101. [CrossRef]
41. Zaman, A.; Obimbo, C.; Dara, R.A. A novel differential privacy approach that enhances classification accuracy. In Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering, Porto, Portugal, 20–22 July 2016; pp. 79–84.
42. Srijayanthi, S.; Sethukarasi, T. Design of privacy preserving model based on clustering involved anonymization along with feature selection. *Comput. Secur.* **2023**, *126*, 103027. [CrossRef]
43. Chen, L.; Zeng, L.; Mu, Y.; Chen, L. Global Combination and Clustering based Differential Privacy Mixed Data Publishing. *IEEE Trans. Knowl. Data Eng.* **2023**. [CrossRef]
44. Jha, N.; Vassio, L.; Trevisan, M.; Leonardi, E.; Mellia, M. Practical anonymization for data streams: Z-anonymity and relation with k-anonymity. *Perform. Eval.* **2023**, *159*, 102329. [CrossRef]
45. Li, B.; He, K. Local generalization and bucketization technique for personalized privacy preservation. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 393–404. [CrossRef]
46. Chu, Z.; He, J.; Li, J.; Wang, Q.; Zhang, X.; Zhu, N. SSKM_DP: Differential Privacy Data Publishing Method via SFLA-Kohonen Network. *Appl. Sci.* **2023**, *13*, 3823. [CrossRef]
47. Sun, X.; Ye, Q.; Hu, H.; Wang, Y.; Huang, K.; Wo, T.; Xu, J. Synthesizing Realistic Trajectory Data With Differential Privacy. *IEEE Trans. Intell. Transp. Syst.* **2023**. [CrossRef]
48. Nóbrega, T.; Pires, C.E.S.; Nascimento, D.C.; Marinho, L.B. Towards automatic Privacy-Preserving Record Linkage: A Transfer Learning based classification step. *Data Knowl. Eng.* **2023**, *145*, 102180. [CrossRef]
49. Amiri, F.; Khan, R.; Anjum, A.; Syed, M.H.; Rehman, S. Enhancing Utility in Anonymized Data against the Adversary's Background Knowledge. *Appl. Sci.* **2023**, *13*, 4091. [CrossRef]
50. Chen, M.; Cang, L.S.; Chang, Z.; Iqbal, M.; Almakhles, D. Data anonymization evaluation against re-identification attacks in edge storage. *Wirel. Netw.* **2023**, 1–15. [CrossRef]
51. Xia, Y.; Zhao, T.; Lv, Y.; Li, Y.; Yang, R. Hierarchical DP-K Anonymous Data Publishing Model Based on Binary Tree. In Proceedings of the 2023 25th International Conference on Advanced Communication Technology (ICACT), Pyeongchang, Republic of Korea, 19–22 February 2023; pp. 102–110.
52. Han, J.; Yu, J.; Lu, J.; Peng, H.; Wu, J. An anonymization method to improve data utility for classification. In Proceedings of the International Symposium on Cyberspace Safety and Security, Xi'an China, 23–25 October 2017; Springer: Cham, Switzerland, 2017; pp. 57–71.
53. Last, M.; Tassa, T.; Zhmudyak, A.; Shmueli, E. Improving accuracy of classification models induced from anonymized datasets. *Inf. Sci.* **2014**, *256*, 138–161. [CrossRef]
54. Fong, P.K.; Weber-Jahnke, J.H. Privacy preserving decision tree learning using unrealized data sets. *IEEE Trans. Knowl. Data Eng.* **2010**, *24*, 353–364. [CrossRef]
55. Lin, K.P.; Chen, M.S. On the design and analysis of the privacy-preserving SVM classifier. *IEEE Trans. Knowl. Data Eng.* **2010**, *23*, 1704–1717. [CrossRef]
56. Park, S.; Byun, J.; Lee, J.; Cheon, J.H.; Lee, J. HE-friendly algorithm for privacy-preserving SVM training. *IEEE Access* **2020**, *8*, 57414–57425. [CrossRef]
57. Eyupoglu, C.; Aydin, M.A.; Zaim, A.H.; Sertbas, A. An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy* **2018**, *20*, 373. [CrossRef]

58. Ye, H.; Chen, E.S. Attribute utility motivated k-anonymization of datasets to support the heterogeneous needs of biomedical researchers. In Proceedings of the AMIA Annual Symposium Proceedings, American Medical Informatics Association, Washington, DC, USA, 22–26 October 2011; Volume 2011, p. 1573.
59. Kousika, N.; Premalatha, K. An improved privacy-preserving data mining technique using singular value decomposition with three-dimensional rotation data perturbation. *J. Supercomput.* **2021**, *77*, 10003–10011. [CrossRef]
60. Selvi, U.; Pushpa, S. Big Data Feature Selection to Achieve Anonymization. In Proceedings of the International Conference on Communication, Computing and Electronics Systems, Coimbatore, India, 21–22 October 2020; Springer: Singapore, 2020; pp. 59–67.
61. Zhang, C.; Jiang, H.; Wang, Y.; Hu, Q.; Yu, J.; Cheng, X. User identity De-anonymization based on attributes. In Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications, Honolulu, HI, USA, 24–26 June 2019; Springer: Cham, Switzerland, 2019; pp. 458–469.
62. Ienca, M.; Vayena, E. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nat. Med.* **2020**, *26*, 463–464. [CrossRef]
63. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
64. MacNell, N.; Feinstein, L.; Wilkerson, J.; Salo, P.M.; Molsberry, S.A.; Fessler, M.B.; Thorne, P.S.; Motsinger-Reif, A.A.; Zeldin, D.C. Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting. *PLoS ONE* **2023**, *18*, e0280387. [CrossRef]
65. Fernández, J.D.; Kirrane, S.; Polleres, A.; Wenning, R. SPECIAL: Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance. 2018. Available online: https://ceur-ws.org/Vol-2044/paper23/paper23.pdf (accessed on 7 April 2023).
66. Gerl, A.; Bennani, N.; Kosch, H.; Brunie, L. LPL, towards a GDPR-compliant privacy language: Formal definition and usage. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXVII*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 41–80.
67. Becher, S.; Gerl, A. ConTra Preference Language: Privacy Preference Unification via Privacy Interfaces. *Sensors* **2022**, *22*, 5428. [CrossRef]
68. Ye, J. Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Math. Comput. Model.* **2011**, *53*, 91–97. [CrossRef]
69. Fkih, F. Similarity Measures for Collaborative Filtering-based Recommender Systems: Review and Experimental Comparison. *J. King Saud Univ.-Comput. Inf. Sci.* **2021**, *34*, 7645–7669. [CrossRef]
70. Liu, J.; Xiong, L.; Luo, J. Semantic Security: Privacy Definitions Revisited. *Trans. Data Priv.* **2013**, *6*, 185–198.
71. Newman, D. UCI Repository of Machine Learning Databases, University of California, Irvine. 1998. Available online: http://www.ics.uci.edu/mlearn/MLRepository.html (accessed on 8 January 2023).
72. Amiri, F.; Yazdani, N.; Shakery, A.; Chinaei, A.H. Hierarchical anonymization algorithms against background knowledge attack in data releasing. *Knowl.-Based Syst.* **2016**, *101*, 71–89. [CrossRef]
73. Fung, B.C.; Wang, K.; Fu, A.W.C.; Philip, S.Y. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*; CRC Press: Boca Raton, FL, USA, 2010.