applied sciences

MDPI

Article

Enhancing Utility in Anonymized Data against the Adversary's Background Knowledge

Fatemeh Amiri [1,*], Razaullah Khan [2,3], Adeel Anjum [4], Madiha Haider Syed [4] and Semeen Rehman [5,*]

[1] Department of Computer Engineering, Hamedan University of Technology, Hamedan 65155, Iran
[2] Department of Computer Science, University of Engineering and Technology, Mardan 23200, Pakistan
[3] National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China
[4] Institute of Information Technology, Quaid-i-Azam University, Islamabad 15320, Pakistan
[5] Institute of Computer Engineering, Technische Universitat Wien (TU Wien), 1040 Wien, Austria
* Correspondence: f.amiri@hut.ac.ir (F.A.); semeen.rehman@tuwien.ac.at (S.R.)

Abstract: Recent studies have shown that data are some of the most valuable resources for making government policies and business decisions in different organizations. In privacy preserving, the challenging task is to keep an individual's data protected and private, and at the same time the modified data must have sufficient accuracy for answering data mining queries. However, it is difficult to implement sufficient privacy where re-identification of a record is claimed to be impossible because the adversary has background knowledge from different sources. The *k*-anonymity model is prone to attribute disclosure, while the *t*-closeness model does not prevent identity disclosure. Moreover, both models do not consider background knowledge attacks. This paper proposes an anonymization algorithm called the utility-based hierarchical algorithm (UHRA) for producing *k*-anonymous *t*-closed data that can prevent background knowledge attacks. The proposed framework satisfies the privacy requirements using a hierarchical approach. Finally, to enhance utility of the anonymized data, records are moved between different anonymized groups, while the requirements of the privacy model are not violated. Our experiments indicate that our proposed algorithm outperforms its counterparts in terms of data utility and privacy.

Keywords: anonymization algorithm; background knowledge attack; data utility; *k*-anonymity; *t*-closeness

check for updates

Citation: 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. https://doi.org/10.3390/app13074091

Academic Editor: Luis Javier García Villalba

Received: 29 January 2023
Revised: 6 March 2023
Accepted: 21 March 2023
Published: 23 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, Big Data and Internet of Things (IoT) technologies have received serious attention because of the advancement in information and communications. Individual data can provide valuable insights to recipients for making business decisions, such as iPhone, Facebook, Twitter, LinkedIn, and some government departments. However, sensitive information, such as personal details about an individual's health, bank accounts, insurance, etc., may be at risk of disclosure if the recipient performs any analysis on the collected data. Therefore, organizations have a responsibility to develop privacy techniques that protect individual identities while disseminating data. Privacy-preserving data publishing (PPDP) is a technique used for disseminating user logs in search engines, graph data, such as social networks, high-dimensional data, such as DNA strings, and patient data at various stages of their diseases while preserving privacy [1–6].

In privacy-preserving data publishing (PPDP), data anonymization is one of the approaches used to protect users' sensitive information [2]. Anonymization refers to a modified version of the original microdata that does not display confidential personal information or any relations between individuals and records in the data. The simplest approach involves removing identity characteristics, such as names and national IDs of individuals [2]. However, this technique has been proven insecure [7]. Samarati [7]

demonstrated that it is possible to identify 87% of US residents using a combination of their date of birth, five-digit zip code, and gender. The challenge of privacy preservation during data publishing involves releasing a modified form of the data that does not reveal identities or sensitive information (i.e., data are no longer personal). It is worth noting that reducing the accuracy of data analysis on anonymous data is inevitable and results in information loss.

Anonymization frameworks typically consist of an anonymization algorithm and a privacy model. Privacy models can be classified into two categories: syntactic and semantic. Syntactic models divide data into several groups, called quasi identifier (QI) groups, such that records in each QI group share the same QI attribute values (Assuming that data are stored in a relational data model, each record has three types of attributes: identifiers, QIs, and sensitive attributes (SA). For instance, in a patient dataset, the name is an identifier attribute, while age, gender, and geographic location are QI attributes, and disease is an SA). The initial syntactic model is k-anonymity [7], where each record must be similar to at least k-1 other records. While this model prevents identity disclosure, it remains vulnerable to attribute disclosure [7]. Several refinements of k-anonymity, such as t-closeness [8] and l-diversity [9], have been proposed to address attribute disclosure. In semantic models, noise is added to attribute values to protect individuals' privacy [10].

Each anonymization framework gives protection against a special adversary model. The adversary model includes assumptions about its knowledge. It is often assumed that (1) the adversary's victims are in the data and (2) the victims' QI attribute values are known [1–9]. Background knowledge (BK) is an additional fact that can help the adversary make accurate inferences on the victim's SA, named a BK attack. This work assumes the adversary's correlational background knowledge (CBK) about the correlations between SA values and QI values. Suppose a hospital publishes patient data anonymously to analyze the progression of diseases (A hospital may publish data on patients for various purposes. For example, it publishes data to analyze the effectiveness of a specific drug in stopping or slowing the disease's progression). CBK in medicine and healthcare can show the impact of age and gender (or individuals' geographical location) on the prevalence of various diseases. Examples of BK in a medical domain are "the prevalence of Bronchitis and Alzheimer's diseases among women 65 years old and older is higher than the prevalence among men in the same age group", "breast cancer in men is rare", etc. [11].

The work in this paper proposes a syntactic anonymization framework that enhances the data utility and prevents attribute disclosure, identity disclosure, and BK attack. When a QI group is released, the adversary can calculate the likelihood of potential connections between the individual and their sensitive information. In the absence of any background knowledge, in cases where there are multiple distinct sensitive values within the QI group, the probability of linking a record respondent to a specific sensitive value is equivalent. Modeling the adversary's BK is an open challenge in PPDP [6,12]. The proposed privacy model aims to minimize certainties in distinguishing record owners and their sensitive values in a QI group. We attempt to produce QI groups where patient records have similar BK distributions to prevent BK attacks. Therefore, adversaries will have reduced confidence in their ability to identify any specific association between a victim and their sensitive values. Ensuring similar BK distributions helps prevent CBK attacks, but it does not provide protection against attribute and identity disclosure. To meet the privacy goals of preventing attribute and identity disclosure, it is crucial to also ensure k-anonymity and t-closeness in each QI group. Therefore, our proposed framework produces QI groups that satisfy three privacy requirements: (1) similar BK distributions, (2) k-anonymity, and (3) t-closeness.

The proposed algorithm uses value generalization. A hierarchical procedure is implemented to fulfill the requirements of the proposed privacy model. Firstly, it uses agglomerative clustering to create clusters with the constraint that the difference of BK distributions between any pair of records within a cluster does not exceed a predefined threshold. Next, the anonymization algorithm partitioned each cluster into QI groups satisfying both t-closeness and k-anonymity requirements. Finally, records within each QI

group are reordered to enhance the similarity of QI values, thereby improving the utility of the anonymized data.

We performed comprehensive experiments using two datasets, namely the adult dataset [13] and the BKseq dataset [6] (Accessed January 2023), to assess the efficacy of our anonymization platform. The evaluation assessed the proposed algorithm against a range of privacy and information loss measures. The results confirm that our algorithm creates anonymous data with high data utility. Moreover, we compared the proposed algorithm with two state-of-the-art anonymization algorithms: the *k*-anonymity-primacy algorithm [12] and the BK-based algorithm [6]. The experimental results reveal that the proposed algorithm outperforms the state-of-the-art anonymization algorithms because of low information loss and privacy loss.

We have made the following main contributions:

1-We propose an anonymization framework that simultaneously satisfies *t*-closeness and *k*-anonymity. Our framework also protects against adversaries with knowledge on correlations of attribute values.

2-We propose a generalization-based anonymization algorithm that satisfies the privacy models and improves data utility. To this end, the records are moved between the QI groups to maximize the usefulness of resultant anonymous data. We conduct comprehensive experiments on various aspects of the algorithms, including the runtime, privacy loss, data utility, and QI groups' size.

The rest of this paper is organized as follows. Section 2 demonstrates the works related to the privacy model, anonymization algorithm, and the adversary's BK. In Section 3, we define the problem, which includes the adversary's BK. Section 4 represents our anonymization framework containing an anonymization algorithm and a privacy model. In Section 5, we present the empirical evaluation, while the conclusion and future work are discussed in Section 6.

2. Related Work

The anonymization techniques are categorized below with respect to the adversary's background knowledge.

2.1. Anonymization without Adversary's Background Knowledge

The *k*-anonymity [7] and its refinements (*t*-closeness [8] and *l*-diversity [9]) are privacy models which divide the records into several QI groups. In *k*-anonymity, the possibility of identifying a record in a QI group is $1/k$ at most. In *t*-closeness, the distance between the distribution of the sensitive values in each QI group and those in the total data is no more than a threshold t value. It should be noted that the refinements (such as *l*-diversity and *t*-closeness) cannot replace the *k*-anonymity approach. Several studies have suggested a combination of privacy mechanisms to prevent against attribute and identity attacks [6,12,14]. Cao and Karras [15] recognized some drawbacks of *t*-closeness that have attracted broad attention to prevent against attribute disclosure [5,12,16–18].

Many different anonymization algorithms have been proposed for satisfying syntactic privacy models. The algorithms use anonymization operations, such as generalization [12,19–24], anatomy [16], and microaggregations [14,17,25]. The generalization refers to altering attribute values in microdata with a range of general values. For example, assume that Table 1 shows patient's data in a hospital. In Table 1, the microdata includes one record for each patient, called a record respondent. Each record in the dataset contains an identifier attribute (name), three quasi-identifier attributes (gender, age, and zip code), and a sensitive attribute (disease). Table 2 shows the corresponding generalized version of the data. An anonymization algorithm that uses generalization deletes identifiers and divides data into QI groups. The SA is published without a change to analyze the data.

Table 1. Original microdata.

Name	Gender	Age	Zip-Code	Disease
Cayla	F	65	12040	Cancer-II
Ganya	F	66	12041	Flu
Elisa	F	66	12041	Depression
Dior	F	66	12041	GRED
Harriet	F	67	12041	Alzheimer-I
Fiona	F	65	12041	Diabetes-II

Table 2. Anonymized data.

QI Group No.	Gender	Age	Zip-Code	Disease
1	F	[65–66]	1204*	Depression
1	F	[65–66]	1204*	GRED
1	F	[65–66]	1204*	Cancer-II
2	F	[65–67]	12041	Alzheimer-I
2	F	[65–67]	12041	Flu
2	F	[65–67]	12041	Diabetes-II

Table 2 satisfies the 3-anonymous model because if the adversary knows the QI values of a victim, without any additional knowledge, the adversary can find the victim's record and disease with a maximum probability of 1/3.

A generalization operation is suitable for categorical data but cannot process numerical data with high accuracy. The anatomy operator results in more than one table and increases the difficulty of data processing and analysis. Microaggregation [17] is not suitable for categorical values. The synthetic records that are released no longer correspond to the actual values in the original dataset, rendering the released data meaningless. In contrast, the generalized data remain semantically compatible with the original raw data. However, due to the elimination of some values, the distribution of anonymous data using microaggregation may differ from that of the original data [12]. This work uses generalization due to its advantages over microaggregation and anatomy. It should be worth noting that each of these operations has its place in research.

A close work to ours is probably the one in [6]. The algorithm creates a list of records according to the Hilbert index. With this technique, records with similar QI values are also close, with high probability. Each k record satisfying t-closeness is released as a QI group. If the condition is not met, other records are added to the QI group. The main problem of said algorithm is high information loss. Another work that is close to ours is the one in [12]. The algorithm satisfies both β–likeness [15] and k-anonymity. They suggested a hierarchical algorithm to meet privacy models. In the algorithm, k-anonymity has priority over β–likeness. The experimental result showed that data utility is low.

2.2. Anonymization with Adversary's Background Knowledge

On the other hand, the type of adversary knowledge is one of the open challenges. None of the above models can preserve privacy against an adversary with BK [6,26,27]. The correlation between attribute values is a significant BK that complicates privacy protection [6,26,28,29]. This work mitigates the risk posed by data correlation between sensitive and quasi-identifier (QI) attributes. Methods for modeling the BK are divided into logic-based [27,30] and probability-based tools [6,18,26,31]. The first category proposes a language for describing a particular level of knowledge and a set of techniques for limiting disclosure attacks. The second category models background knowledge (BK) as a probability distribution that links sensitive attribute values to individual respondents in the dataset. Another type of privacy models are the differential privacy models. These models have also attracted attention because of some noise addition and are categorized as semantic privacy models [32,33]. However, studies have shown that such models are ineffective while the adversary possesses some BK [34]. Therefore, the work focused in this research (syntactic approach such as k-anonymity and its refinements) has more utility and

privacy compared to other privacy measures. In addition, the proposed work is aware of any possible attack via the adversary's BK.

3. Problem Definition

Assume a data publisher plans to release the table T = {r_1, r_2, …, r_n}, where every record r_i belongs to an individual v_i and includes an identifier ID, QI attributes A_1,A_2, …, A_D, and one SI of A_{D+1}. Here, [A_j] shows the domain of A_j for 1≤ j ≤ D + 1; $r_i(A_j)$ is the value for attribute A_j in record r_i, and r_i(QI) denotes the values for all QI attributes of record r_i. This work assumes that: (1) the adversary is aware of the existence of the victims and their QI values, (2) the adversary has BK regarding the correlation between SA values and QI values which is described in Section 3.1.

3.1. Background Knowledge

In this work, we adopt the concept of background knowledge (BK) to represent the prior probability of accurately inferring sensitive attribute values based on quasi-identifier (QI) attributes for a given record respondent. The adversary's BK is defined as a function PD^{sv}:$[QI] \rightarrow \sum$, where $[QI] = [A_1]\times\ldots\times[A_D]$ is the set of all possible quasi-identifier values, and $\sum = \{(p_1, p_2, \ldots, p_M) \,|p_i \in \mathbb{R}, \sum_{i=1}^{M} p_i = 1, 0 \leq p_i \leq 1\}$ is the set of all possible probability distributions, and M is the number of different sensitive values.

Therefore, for a record respondent v with QI values of q, PD^{sv} is modeled as (p_1, p_2, …, p_M) over the domain of SA, where p_i is the probability of assigning v to the sensitive value s_i given q. Therefore, each record in table T has a BK distribution. BK is distinct from the distribution of sensitive values, which indicates the frequency of their occurrence in dataset T. There are practical algorithms to obtain BK according to the available external data. This study used the incremental method [6] to extract the BK distributions. The mathematical symbols used in this study are provided in Appendix A.

4. Our Anonymization Framework

A syntactic privacy model aims to minimize the certainty in identifying each record's actual respondent in a given QI group. To this end, the anonymization framework reduces the adversary's confidence in separating one association among all possible associations. Based on the knowledge definition in Section 3.1, our anonymization algorithm creates QI groups where record respondents' BK distributions are the same.

Lemma 1. *Assume that Σ_Q is a set of BK distributions of record owners in QI group q. The equal probability distributions in q minimizes the certainty in identifying the true record owners in q.*

Lemma 1 was proven by Amiri et al. [12]. Since it is challenging to create QI groups with similar BK distributions, We enforce a constraint that limits the divergence between these distributions within a quasi-identifier (QI) group to be below a threshold J. Since the condition is not sufficient to prevent attribute and identity disclosure risks, this work also uses *t*-closeness [8] and *k*-anonymity [7]. Therefore, our privacy requirements are defined as the following:

(1) The difference of BK distributions in every QI group is up to the threshold J;
(2) The *k*-anonymity model;
(3) The *t*-closeness model.

This work suggests the utility-based hierarchical algorithm (UHRA), which hierarchically meets the privacy requirements. As can be observed in Algorithm 1, the proposed algorithm includes two steps:

(1) BK-based clustering (line 2 of Algorithm 1): It partitions the table into a few clusters {N_1, …, $N_{|N|}$} so that in each cluster the difference in BK distributions is equal to or less than J;

(2) The *t*-*k*-Utility (line 5 in Algorithm 1): To fulfill *t*-closeness and *k*-anonymity, each cluster N_i is partitioned into a few QI groups.

Algorithm 1: UHRA Algorithm.

Input:
$T = \{r_1, r_2, \ldots, r_n\}$: The original microdata;
PD^{sv}: Background Knowledge distributions of table T;
J, t, *and* k: the parameters in privacy model.
Output:
A set of QI groups $\{q_1, \ldots, q_h\}$

1 *//Step 1: To generate a set of clusters* $N = \{N_1, \ldots, N_{|N|}\}$
2 **Background knowledge-based clustering**
3 **for** each cluster N_i such that $|N_i| \geq$k **do**
4 *//Step 2: To partition records within* N_i *into QI groups*
5 ***t-k*-Utility**
6 **end for**

In the first step, a cluster of records is created using hierarchical agglomerative clustering. Jensen Shannon Divergence (JSD) [35] is used to calculate the similarity between probability distributions. The cutoff point is defined as the threshold J, below which the distribution difference between a node and its descendants is considered small enough. This step utilizes the complete linkage method to guarantee that the distance between any two records in a cluster is no greater than J. This step outputs a set of clusters, $\{N_1, \ldots, N_{|N|}\}$, and for each cluster N_i with at least k records, the *t*-*k*-Utility algorithm is executed. Clusters with less than k records are not disseminated.

The *t*-*k*-Utility algorithm partitions the cluster N_i into groups where k-anonymity and t-closeness are fulfilled. To enhance utility, records are moved between groups without violating the privacy requirements. The following lemma describes a guarantee condition in which the *t*-*k*-Utility produces k-anonymous data satisfying t-closeness.

Lemma 2. *Assume* N_i *is a record cluster* ($N_i \in N$, $N_i \subseteq T$, *and* $|N_i| \geq k$), *and EMD*(N_i,T) *(Earth Mover's Distance [36]) calculates the difference between the distribution of* N_i*'s sensitive values and that in* T. *The t-k-Utility outputs k-anonymous t-closeness data if EMD*(N_i,T) $\leq t$.

Proof. In the worst case, *t*-*k*-Utility produces a QI group consisting of complete microdata in N_i ($N_i \geq k$). When EMD(N_i,T) $\leq t$, N_i satisfies t-closeness and k-anonymity. Hence, the *t*-*k*-Utility outputs a t-closeness k-anonymous dataset. □

Pseudo-code of *t*-*k*-Utility is shown in Algorithm 2. If N_i is a set of records, P_T is the distribution of SA values in T, and $[A_{D+1}]_{N_i}$ is the domain of SAs in N_i. Then, *t*-*k*-Utility accepts N_i, P_T, $[A_{D+1}]_{N_i}$, k, and t as inputs and creates a set of QI groups. The proposed *t*-*k*-Utility consists of two steps: the creation of initial QI groups (Line 2 to 20 of Algorithm 2) and the refinement of QI groups (Line 22 to 28 of Algorithm 2). In the first step, records are partitioned into groups that satisfy t-closeness and k-anonymity conditions. In the second step, records are moved between different groups to improve the data utility. In Section 4.1, we outline the process for creating primary QI groups, while in Section 4.2, we detail the steps for refining the final QI groups.

4.1. Creation of the Initial QI Groups

At first, the *t*-*k*-Utility algorithm generates a sorted list, π, of complete records in N_i (Line 3 in Algorithm 2). The *t*-*k*-Utility uses the nearest point next (NPN) technique [9] to create the list (After evaluating various techniques to create an ordered list of records based on QI values, including space-filling curves [6], we found that the experimental results were consistent for both the adult dataset and the BKseq dataset). It sorts the records according to the QI values.

The *t*-*k*-Utility algorithm follows a two-step process to ensure privacy and limit distortion caused by generalizing QI values. First, it creates primary QI groups by selecting the first k records as group q and checks if t-closeness is met. If not, it adds the next record to q and repeats the process until t-closeness is achieved (Lines 6 to 12 in Algorithm 2). Second,

the algorithm attempts to add each record to an existing QI group, q′, while maintaining *t*-closeness, and removes any records that cannot be placed in a group (Lines 13 to 21 in Algorithm 2). If *t*-closeness cannot be achieved, other solutions may be considered [6].

Algorithm 2: The *t*-*k*-Utility algorithm.

Input:
N_i: The set of records;
$[A_{D+1}]_{N_i}$: Domain of sensitive value in N_i;
P_T: Distribution of sensitive values in T;
k and t: The parameters in privacy model
Output:
A set of QI groups satisfying privacy model

1 *// Step 1: To Create initial QI groups of N_i as $\{q_1,q_2,\ldots,q_c\}$*
2 $Qtemp \leftarrow \emptyset$ *// a set of primary QI groups*
3 $\pi \leftarrow$ Create an order of all $|N_i|$ records using nearest point next method
4 $q \leftarrow \emptyset$ *// an empty QI group*
5 **for** $\tilde{r} \leftarrow \pi_1$ to $\pi_{|N_i|}$ **do**
6 $q \leftarrow q \cup \{\tilde{r}\}$
7 **if** $|q| \geq k$ **and** t-closeness(q) **then**
8 $Qtemp \leftarrow Qtemp \cup \{q\}$
9 $q \leftarrow \emptyset$
10 **end if**
11 **end for**
12 **if** $q \neq \emptyset$ **then**
13 **for** $r \in q$ **do**
14 **if** $\exists\, q' \in Qtemp$ such that t-closeness($\{q' \cup \{r\}\}$) **then**
15 $Qtemp \leftarrow (Qtemp \cup \{q' \cup \{r\}\}) \setminus q'$ ($x \backslash \{r\}$ means all receords in set x except the receord r)
16 **else**
17 Remove r
18 **end if**
19 **end for**
20 **end if**
21 *// Step 2: To refine the QI groups $\{q_1,q_2,q_3,\ldots,q_c\}$*
22 **Repeat**
23 **for** each $r \in N_i$ **do**
24 Compute the distance between r and the QI groups in $\{q_1,q_2,q_3,\ldots,q_c\}$
25 Assign r to the closest QI group
26 **end for**
27 **Until** the QI groups does not change
28 **end function**

4.2. Refinement of the QI Groups

In the second step, QI groups are iteratively refined by calculating the distance between records and QI groups. If the privacy model is not violated, each record is re-assigned to the closest QI group (with closest QI values). Refinement is continued until no change is created in the groups. Finally, each group is disseminated as a QI group. We must avoid removing (or adding) a record from a QI group because if we do so, it violates the *k*-anonymity and *t*-closeness techniques. The refinement is achieved by defining the distance $d'(r, q_i)$ between the record r and the QI group q_i (line 25 in Algorithm 2) as:

$$d'(r,q_i) = \begin{cases} \infty & \text{if } r \notin q_i \text{ And EMD}(q_i \cup \{r\},T) > t \\ 0 & \text{if } r \in q_i \text{ And (EMD}(q_i \backslash \{r\},T) > t \text{ Or } |q_i \backslash \{r\}| < k) \\ d(r,q_i) & \text{otherwise} \end{cases} \quad (1)$$

where $|q_i \backslash \{r\}|$ shows the number of records in $q_i \backslash \{r\}$. The first and second cases in Equation (1) prevent adding and removing records if it violates the *k*-anonymity and *t*-closeness. The third case calculates the Euclidean distance between r and q_i according to their QI values.

4.3. Efficiency of the Proposed Anonymization Algorithm

The proposed anonymization algorithm includes two steps: BK-based clustering and the *t*-*k*-Utility approach. BK-based clustering applies agglomerative clustering and outputs a set of clusters, $N = \{N_1, \ldots, N_{|N|}\}$. The analysis of the algorithm shows that agglomerative clustering is a time-consuming component. Its worst-case time complexity is $O(n^2 logn)$ when n is the number of records in the original microdata. Then, we should apply *t*-*k*-Utility for each cluster $N_i \in N$. In the first step of *t*-*k*-Utility, the NPN method has an $O(|N_i|^2)$ time complexity, and the worst-case running time of the second step of *t*-*k*-Utility is also $O(|N_i|^2)$. Hence, the time complexity of the proposed algorithm is $O(n^2 logn + \sum_{N_i \in N} O(|N_i|^2))$. In the worst case, BK-based clustering creates a cluster with the size n. So, the worst-case time complexity of the proposed algorithm will be $O(n^2 logn + n^2)$.

5. Experimental Result

In the following section, we will conduct an empirical evaluation of the proposed algorithm using two different datasets and various analysis measures. Section 5.1 explains the experiments' configuration and datasets, and Section 5.2 presents the evaluation criteria. Finally, Section 5.3 shows the evaluation results for the proposed algorithm.

5.1. Evaluation Datasets and Configuration

To assess the efficacy of our proposed algorithms, we conducted several experiments using a machine equipped with a 2.5 GHz Intel Core i7 processor and 8GB RAM. For implementing the hierarchical dendrogram, we utilized the linkage function provided by MATLAB (version R2022a). The distance between two clusters is measured using the complete linkage method. We conducted experiments with various values of privacy parameters (J, *k*, and *t*) to evaluate the effectiveness of our proposed algorithms. The values for the model parameters were selected based on the policies of the data domain. To compare our algorithm with other similar studies [6,12], we took the value of J, *k*, and *t* in the ranges of 0.2–0.8, 3–20, and 0.5–0.8. It should be noted that each experiment was executed ten times, and the average of results is reported for smooth graph display. Experiments were run on two datasets consisting of categorical and numerical attributes as follows:

(1) The adult dataset [13] is a dataset published by the US statistics center in 1994, which is widely used for anonymization experiments. This dataset contains 45,222 statistical records with 8 grouping attributes, 6 numerical attributes, and 1 attribute for showing income. To compare our algorithm with other similar studies [6,12], three attributes of age, gender, and education were used as QI attributes, while income was used as an SA.

(2) The BKseq dataset [6] was generated based on domain knowledge extracted from medical literature. It includes a history of 24 tables, each table including 4000 records. Each record in the dataset comprises three quasi-identifier (QI) attributes: gender, age, and weight, as well as a sensitive attribute that reflects the result of medical exams. The sensitive attribute includes 19 distinct values representing various stages of the disease.

5.2. Evaluation Measure

We applied global certainty penalty (GCP) to evaluate the impact on data utility generated by anonymization (Different data utility measures are suggested for anonymization algorithms. GCP is a well-known measure and does not have limitations that other measures do [12]). GCP uses the normalized certainty penalty (*NCP*) measure. If table T includes D numerical QI attributes, $A_1, A_2, \ldots, A_D$; the generalized version of a record r = $(x_1, \ldots, x_D)$ is $r' = ([y_1 - z_1], \ldots, [y_D - z_D])$ so that $\forall 1 \leq i \leq D, y_i \leq z_i$ (x_i, y_i and z_i are numbers). *NCP* on A_i is calculated in Equation (2).

$$NCP_{A_i}(r) = \frac{z_i - y_i}{|A_i|} \tag{2}$$

where the expression $|A_i|$ represents the number of unique values that attribute A_i takes in table T. Based on this, the $NCP(r)$ is computed as a weighted sum of its NCP values across all QI attributes, as given by Equation (3).

$$NCP(r) = \sum_{i=1}^{n} (w_i \times NCP_{A_i}(r)) \tag{3}$$

GCP is calculated as the sum of NCP on all records in T. The value for GCP is a number larger or equal to zero, where zero indicates a situation with no information loss and maximum data usefulness. Therefore, lower GCP values are desirable.

We also evaluated our algorithms regarding record linkage (RL) [19], which measures privacy loss. It is used for disclosure risk analysis. RL is calculated using Equation (4) by the number of correct linkages between the original data and the anonymized data.

$$RL = \frac{\sum_{r \in T} P_{RL}(r')}{n} \tag{4}$$

where n is the size of T, and $P_{RL}(r')$ is the linkage probability of the anonymized record r'. The linkage probability of r'_j is calculated via Equation (5) below.

$$P_{RL}(r') = \begin{cases} \frac{1}{|E_i|} & : r' \in E_i \\ 0 & : otherwise \end{cases} \tag{5}$$

where E_i is the closest QI group to r. Here, $r' \in E_i$, $P_{RL}(r')$ is computed as the inverse of the size of E_i; otherwise, the probability is zero. The lower RL means the higher privacy of the record respondents.

5.3. Effectiveness of the Proposed Algorithm

In this section, the behavior of the proposed anonymization algorithm is analyzed from three aspects: the R-U confidentiality map, the size of QI groups, and the runtime.

The R-U confidentiality map is applied as a graphical description of GCP and RL. The performance of the proposed anonymization algorithm UHRA was compared with two other approaches: the BK-based algorithm [6] and k-anonymity primacy [12], shown as BKA-JS and k-primacy in the following tables and figures, respectively. BKA-JS and k-primacy are generalization-based algorithms. We ran their public versions to compare the results. We examine the impact of different privacy parameters: t, k, and J.

5.3.1. R-U Confidentiality Map

We investigated how varying the privacy parameters affects RL and GCP generated by three different algorithms, UHRA, BKA-JS, and k-primacy. First, we assumed that t and J were set to 0.5 and 0.8. The evaluation results of 3 algorithms using adult and BKseq datasets are shown in Figure 1, when k varies from 3 to 20.

BKA-JS creates a sorted list of records based on the Hilbert index of QI values. The algorithm selects the first k records in the list and checks if the JSD values and k-closeness are below thresholds J and t. Then, k-primacy generates groups where the difference between BK distributions does not exceed a threshold J. All three algorithms try to maximize the homogeneity of QI values in QI groups. As seen in Figure 1a, the BKA-JS algorithm results in higher information and privacy losses. The experimental results display that UHRA outperforms BKA-JS and k-primacy due to lower information loss. This result is because UHRA inserts one record into the existing QI groups until t-closeness is met. It then refines the initial QI groups to satisfy the privacy model with homogenous QI values. Our proposed algorithm has the best results with regard to privacy loss too. The results of evaluating the algorithms' performance on the BKseq dataset shown in Figure 1b confirm the findings. Subsequently, we evaluated the algorithms' performance across multiple values of J, with the parameters t and k fixed at 0.5 and 5, respectively. The value of the J

parameter changes in the range of 0.2–0.8. The experimental results presented in Figure 2 restated our previous findings. UHRA results in QI groups with lower RL in both datasets, while the BKA-JS algorithm results in high RL and high privacy loss.

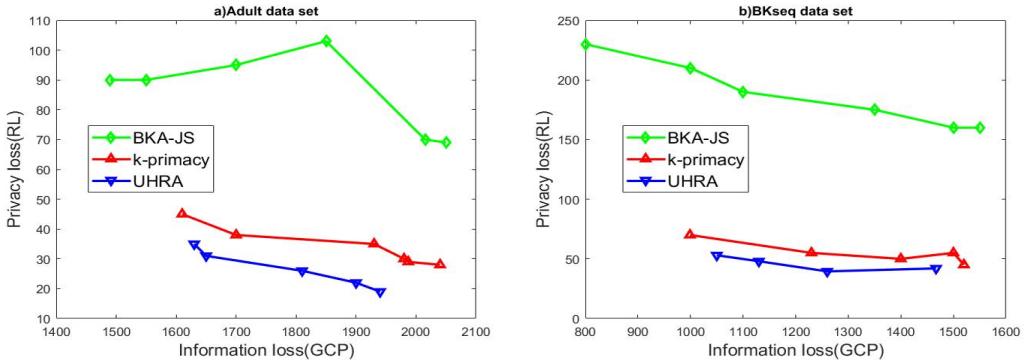


Figure 1. R-U Map regarding various *k* values.

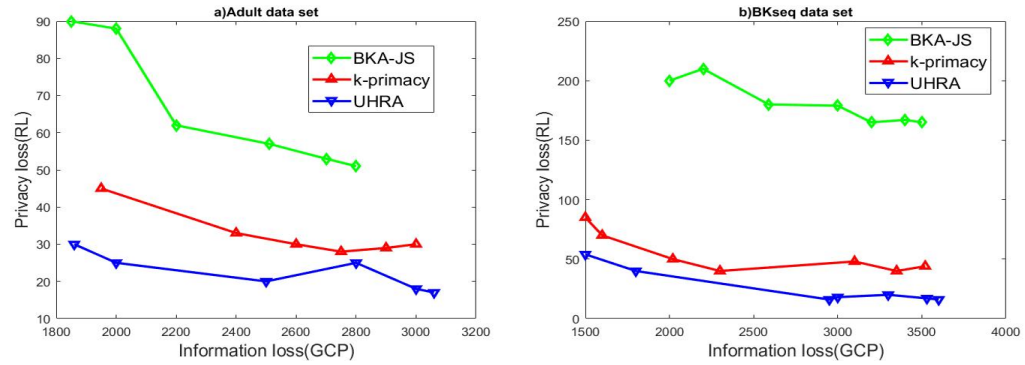


Figure 2. R-U map with regards to various J values.

In the next set of experiments, we analyzed the behavior of our algorithm as *k* and J were fixed. In the analysis, the *t* values vary in the range 0.5–0.8 and J and *k* were set to 0.8 and 5. As seen in Figure 3, The experimental results on both datasets affirm previous findings.

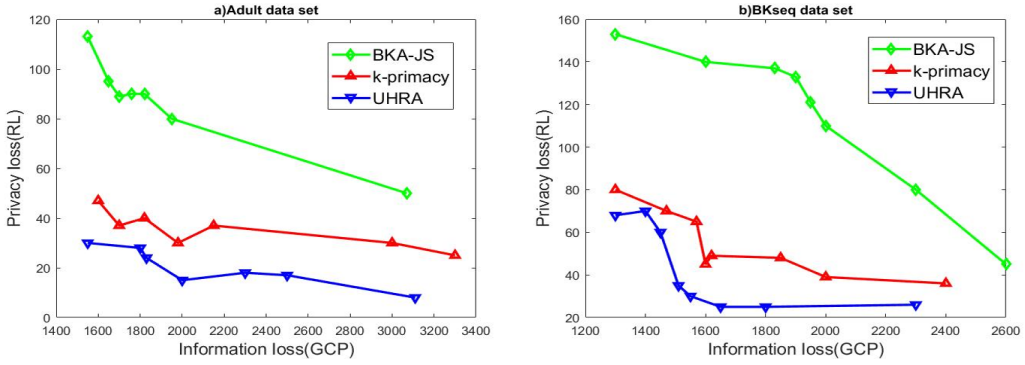


Figure 3. R-U map regarding various *t* values.

Therefore, it can be concluded that the UHRA algorithm results in less information loss and better privacy in its output. In each level of RL, both BKA-JS and *k*-primacy result in higher information loss. Although both *k*-primacy and BKA-JS try to take records with similar QI values to enhance data utility, UHRA outperforms them regarding privacy and information loss. The refinement of the QI groups in UHRA results in lower information loss and higher data utility.

5.3.2. Actual QI Group Size

This section presents how close to k the sizes of QI groups created by algorithms on the Adult dataset are. To minimize information loss and enhance data utility, QI groups with sizes close to k are better. At first, we assume that J and t are constant. The actual size of QI groups is shown in Table 3 over different values of k. Each row presents the minimum and the average size of QI groups over different values of k. The minimum size of QI groups determines the level of k-anonymity. As seen in Table 3, an increase in k values will increase the average size of QI groups. In addition, our algorithm's average size of QI groups is less than BKA-JS and k-primacy. Moreover, the difference between the average and the minimum size of QI groups in BKA-JS is more than other algorithms. We conclude that our proposed algorithm creates smaller QI groups compared to BKA-JS and k-primacy.

Table 3. The size (min, avg) of QI groups on the adult dataset with t = 0.5 and J = 0.8.

Method	k = 3		k = 5		k = 10		k = 15		k = 20	
	min	avg	min	avg	min	avg	min	avg	min	avg
UHRA	**3**	**8**	**5**	**9**	**10**	**11**	**15**	**17**	**20**	**23**
BKA-JS	3	8.8	5	13	10	14	15	20	20	25
k-primacy	3	8	5	11	10	14	15	18	20	24

We then investigate the size of quasi-identifier (QI) groups generated by the algorithms in Table 4, while holding the parameters k and t constant. Each row in the table reports the minimum and average size of QI groups for various J values. The findings confirm earlier observations that increasing J leads to larger QI groups on average. Specifically, we note that the average size of QI groups produced by BKA-JS exceeds that of UHRA.

Table 4. The size (min, avg) of QI groups on Adult dataset with t = 0.5 and k = 5.

Method	J = 0.2		J = 0.3		J = 0.4		J = 0.5		J = 0.6	
	min	avg	min	avg	min	avg	min	avg	min	avg
UHRA	**5**	**9**	**5**	**9**	**5**	**10**	**5**	**11**	**5**	**10**
BKA-JS	5	11	5	11.8	5	13	5	13	5	13
k-primacy	5	10	5	11	5	12	5	12	5	12

Finally, we show how close to k are the sizes of the QI groups produced by algorithms over different values of t. We assume that k and J are set to 5 and 0.8. The actual size of QI groups created by algorithms is presented in Table 5. Each table row shows the minimum and the average size of the QI groups when the t value varies from 0.5 to 0.8. The experimental results also confirm previous conclusions. We see that the average size of QI groups increases for different t values.

Table 5. The size of QI groups on Adult dataset with J = 0.8 and k = 5.

Method	t = 0.5		t = 0.6		t = 0.7		t = 0.8	
	min	avg	min	avg	min	avg	min	avg
UHRA	**5**	**9**	**5**	**10**	**5**	**11**	**5**	**11**
BKA-JS	5	13	5	13	5	15	5	17
k-primacy	5	11	5	12	5	12	5	12.8

5.3.3. The Speed and Scalability

In this section, we analyze the runtime of the proposed algorithm on the BKseq dataset. To this end, k and t are set to 5 and 0.5, while J is taken in the range of 0.1–0.7. The runtime of algorithms for different J values is presented in Figure 4. In the figure, the vertical axis is on a logarithmic scale. Our experiments reveal that the runtime of our algorithms remains consistent for J values greater than or equal to 0.3. This result stems from the close

proximity of background knowledge (BK) distributions among the records in the BKseq dataset and the first step of UHRA generating a single cluster for J values less than or equal to 0.3. Notably, our algorithm exhibits faster execution times for J values less than 0.3. Specifically, UHRA's initial step generates multiple small clusters for these J values, which results in fewer record movements between QI groups to satisfy the privacy constraints. We also see that the runtime of UHRA is higher than BKA-JS, but our algorithm is similar to *k*-primacy.

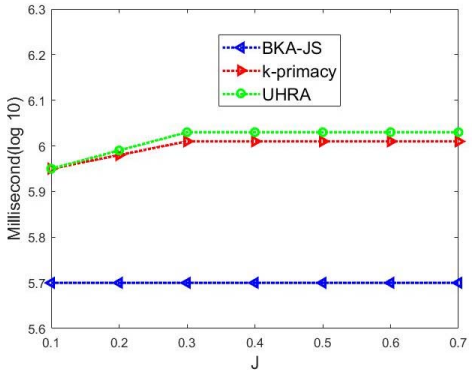


Figure 4. Execution time (in milliseconds with log 10 scale) of our algorithm with t= 0.5 and k = 5 regarding various J values on the BKseq dataset.

Then, we select values of privacy parameters such that a worst-case execution time can be seen. We set J = 0.8, so the BK-based clustering creates clusters with maximum sizes, and *t*-*k*-Utility moves more of the records among QI groups to meet the privacy requirements. The execution time of algorithms on the BKseq dataset is displayed in Figure 5 when the k value varies from 3 to 20. Our algorithm and *k*-primacy have time complexity with the order o(n^2 $logn$) to create the groups in BK-based clustering. We see that the runtime decreases as k increases. It can be seen that BKA-JS is more efficient than our proposed algorithm due to a lower runtime. UHRA and *k*-primacy have a similar runtime.

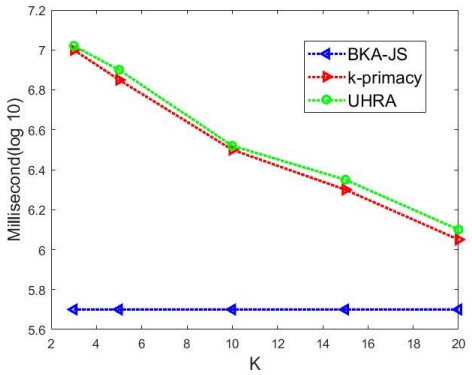


Figure 5. Execution time (in milliseconds with log 10 scale) of our algorithm when J = 0.8, t = 0.5, and the k value varies on the BKseq dataset

6. Conclusions

This study proposed an algorithm for data publishing and dissemination in which the adversary has the BK to re-identify an individual in the anonymized data. However, privacy requirements prevent the BK attacks and identity and attribute disclosures. A hierarchical algorithm, UHRA, was proposed to satisfy the privacy requirements, including two steps. First, UHRA uses agglomerative clustering to prevent background knowledge attacks that consider correlations between sensitive and QI attributes. Then, *t*-closeness and *k*-anonymity are applied to the QI groups, and records are moved between QI groups

to increase data utility as long as the privacy requirements are not violated. The evaluation results confirm the higher data utility compared to other algorithms. Experimental results show that the proposed algorithm causes lower privacy loss and information loss. An increase in the k value increases the size of the QI groups, and the information loss rises too. When the J and t values decrease, the number of unpublished records increases.

For future studies, we suggest that the proposed algorithm be evaluated in re-publishing scenarios. In data re-publishing, an adversary can use the combination of information from different releases over time. Data correlation can lead to a privacy breach.

Author Contributions: Conceptualization, F.A. and M.H.S.; Methodology, F.A.; Formal analysis, F.A and S.R.; Investigation, F.A. and Razaullah Khan; Data curation, F.A., A.A. and S.R.; Writing—original draft, F.A.; Writing—review & editing, F.A., Razaullah Khan and M.H.S.; Supervision, A.A.; Funding acquisition, S.R. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets are used in this study: Adult Dataset. 2022. Available online: https://archive.ics.uci.edu/ml/datasets/Adult (accessed on 22 August 2022).

Acknowledgments: The authors acknowledge TU Wien Bibliothek for supporting this research.

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

Appendix A. Summary of Notations

T	the original table
r_n	the nth record in ith view
r'_n	anonymized version of r_n
A_j	the jth attribute
$[A_j]$	domain of A_j
M	the number of distinct sensitive attribute value
D	the number of QI attributes
PD^{sv}	background knowledge

References

1. Gardner, J.; Xiong, L. An integrated framework for de-identifying unstructured medical data. *Data Knowl. Eng.* **2009**, *68*, 1441–1451. [CrossRef]
2. Fung, B.C.M.; Wang, K.; Fu, A.W.C.; Yu, P. *Introduction to Privacy-Peserving Data Publishing: Concepts and Techniques*; Chapman & Hall/CRC Data Mining and Knowledge Discovery Series; CRC Press: Boca Raton, FL, USA, 2012.
3. Rafiei, M.; van der Aalst, W.M. Group-based privacy preservation techniques for process mining. *Data Knowl. Eng.* **2021**, *134*, 101908. [CrossRef]
4. Al-Hussaeni, K.H.; Fung, B.C.M.; Cheung, W. Privacy-preserving trajectory stream publishing. *Data Knowl. Eng.* **2014**, *94*, 89–109. [CrossRef]
5. Hu, Z.; Yang, J.; Zhang, J. Trajectory privacy protection method based on the time interval divided. *Comput. Secur.* **2018**, *77*, 488–499. [CrossRef]
6. Riboni, D.; Pareschi, L.; Bettini, C. JS-Reduce: Defending Your Data From Sequential Background Knowledge Attacks. *IEEE Trans. Dep. Sec. Comp.* **2012**, *9*, 387–400. [CrossRef]
7. Samarati, P. Protecting Respondents' Identities in Microdata Release. *IEEE Trans. Knowl. Data Eng.* **2001**, *13*, 1010–1027. [CrossRef]
8. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity And L-Diversity. In Proceedings of the 23th IEEE International Conference on Data Eng (ICDE), Istanbul, Turkey, 15–20 April 2007; pp. 106–115.
9. Machanavajjhala, P.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. L-diversity: Privacy Beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3-es. [CrossRef]
10. Dwork, C. Differential Privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), Venice, Italy, 10–14 July 2006; pp. 1–12.
11. National Heart, Lung and Blood Institute, Data Fact Sheet. 2022. Available online: https://apsfa.org/docs/copd_fact.pdf (accessed on 22 August 2022).

12. Amiri, F.; Yazdani, N.; Shakery, A.; Chinaei, A.H. Hierarchical Anonymization Algorithms against Background Knowledge Attack in Data Releasing. *Knowl. Based Sys.* **2016**, *101*, 71–89. [CrossRef]
13. Adult Dataset. 2022. Available online: https://archive.ics.uci.edu/ml/datasets/Adult (accessed on 22 August 2022).
14. Domingo-Ferrer, J.; Soria-Comas, J.; Mulero-Vellido, R. Steered Microaggregation As A Unified Primitive to Anonymize Data Sets and Data Streams. *IEEE Trans. Knowl. Data Eng.* **2019**, *14*, 3298–3311. [CrossRef]
15. Cao, J.; Karras, P. Publishing microdata with a robust privacy guarantee. *Proc. VLDB Endow.* **2012**, *5*, 1388–1399. [CrossRef]
16. Xiao, X.; Tao, Y. Anatomy: Simple and Effective Privacy Preservation. In Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Republic of Korea, 12–15 September 2006; pp. 139–150.
17. Soria-Comas, J.; Domingo-Ferrer, J.; Sánchez, D.; Martínez, S. T-closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 3098–3110. [CrossRef]
18. Amiri, F.; Yazdani, N.; Shakery, A. Bottom-up sequential anonymization in the presence of adversary knowledge. *Inf. Sci.* **2018**, *450*, 316–335. [CrossRef]
19. Li, T.; Li, N.; Zhang, J.; Molloy, I. Slicing: A New Approach for Privacy Preserving Data Publishing. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 561–574. [CrossRef]
20. LeFevre, K.; DeWitt, D.J.; Raghu, R. Mondrian Multidimensional k-Anonymity. In Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), Washington, DC, USA, 3–7 April 2006; pp. 25–35.
21. Li, J.; Liu, J.; Baig, M.; Chi-Wing Wong, R. Information based data anonymization for classification utility. *Data Knowl. Eng.* **2011**, *70*, 1030–1045. [CrossRef]
22. Ercan Nergiz, M.; Clifton, C. Thoughts on k-anonymization. *Data Knowl. Eng.* **2007**, *63*, 622–645. [CrossRef]
23. Can, O. Personalised anonymity for microdata release. *IET Inf. Secur.* **2018**, *2*, 341–347. [CrossRef]
24. Song, H.; Wang, N.; Sun, J.; Luo, T.; Li, J. Enhanced anonymous models for microdata release based on sensitive levels partition. *Comput. Commun.* **2020**, *155*, 9–23. [CrossRef]
25. Soria-Comas, J.; Domingo-Ferrer, J.; Mulero, R. Efficient Near optimal Variable-size Microaggregation. In Proceedings of the Modeling Decisions for Artificial Intelligence—MDAI, Milan, Italy, 4–6 September 2019; pp. 333–345.
26. Majeed, A.; Hwang, S.O. Quantifying the Vulnerability of Attributes for Effective Privacy Preservation Using Machine Learning. *IEEE Access.* **2023**, *11*, 4400–4411. [CrossRef]
27. Li, T.; Li, N. Injector: Mining Background Knowledge For Data Anonymization. In Proceedings of the International Conference on Data Engineering (ICDE), Cancun, Mexico, 7–12 April 2008; pp. 446–455.
28. Wang, H.; Liu, R. Privacy-Preserving Publishing Microdata with Full Functional Dependencies. *Data Knowl. Eng.* **2011**, *70*, 249–268. [CrossRef]
29. Al Bouna, B.; Clifton, C.; Malluhi, Q. Efficient Sanitization of Unsafe Data Correlations. In Proceedings of the Workshops of the EDBT/ICDT 2015 Joint Conference, Brussels, Belgium, 27 March 2015; pp. 278–285.
30. Martin, D.; Kifer, D.; Machanavajjhala, A.; Gehrke, J.; Halpern, J. Worst-case Background Knowledge for Privacy-Preserving Data Publishing. In Proceedings of the International Conference on Data Engineering (ICDE), Istanbul, Turkey, 15–20 April 2007; pp. 126–135.
31. Amiri, F.; Yazdani, N.; Shakery, A.; Ho, S.S. Bayesian-based Anonymization Framework Against Background Knowledge Attack in Continuous Data Publishing. *Trans. Data Priv.* **2019**, *12*, 197–225.
32. Soria-Comas, J.; Domingo-Ferrer, J. Differentially Private Data Publishing via Optimal Univariate Microaggregation and Record Perturbation. *Knowl.-Based Syst.* **2018**, *153*, 78–90. [CrossRef]
33. Zhao, X.; Pi, D.; Chen, J. Novel Trajectory Privacy-preserving Method Based on Prefix Tree Using Differential Privacy. *Knowl.-Based Syst.* **2020**, *198*, 105940. [CrossRef]
34. Kifer, D.; Machanavajjhala, A. No Free Lunch in Data Privacy. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, Athens, Greece, 12–16 June 2011; pp. 193–204.
35. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
36. Rubner, Y.; Tomasi, C.; Guibas, L.J. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.