applied sciences

MDPI

Article

Efficient Multi-Source Anonymity for Aggregated Internet of Vehicles Datasets

Xingmin Lu [1] and Wei Song [2,*]

[1] School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; lxm.xupt@gmail.com

[2] School of Information Science and Technology, North China University of Technology, Beijing 100144, China

* Correspondence: songwei@ncut.edu.cn

Abstract: The widespread use of data makes privacy protection an urgent problem that must be addressed. Anonymity is a traditional technique that is used to protect private information. In multi-source data scenarios, if attackers have background knowledge of the data from one source, they may obtain accurate quasi-identifier (QI) values for other data sources. By analyzing the aggregated dataset, *k*-anonymity generalizes all or part of the QI values. Hence, some values remain unchanged. This creates new privacy disclosures for inferring other information about an individual. However, current techniques cannot address this problem. This study explores the additional privacy disclosures of aggregated datasets. We propose a new attack called a multi-source linkability attack. Subsequently, we design multi-source (*k*,*d*)-anonymity and multi-source (*k*,*l*,*d*)-diversity models and algorithms to protect the quasi-identifiers and sensitive attributes, respectively. We experimentally evaluate our algorithms on real datasets: that is, the Adult and Census datasets. Our work can better prevent privacy disclosures in multi-source scenarios compared to existing Incognito, Flash, Top-down, and Mondrian algorithms. The experimental results also demonstrate that our algorithms perform well regarding information loss and efficiency.

Keywords: multi-source (*k*,*d*)-anonymity; multi-source (*k*,*l*,*d*)-diversity; privacy; IoV; *k*-anonymity; aggregated dataset

check for updates

Citation: Lu, X.; Song, W. Efficient Multi-Source Anonymity for Aggregated Internet of Vehicles Datasets. *Appl. Sci.* **2024**, *14*, 3230. https://doi.org/10.3390/app14083230

Academic Editor: David Megías

Received: 26 January 2024
Revised: 6 April 2024
Accepted: 10 April 2024
Published: 11 April 2024

1. Introduction

In recent years, with the rise and advancement of the Internet, numerous organizations have made substantial amounts of data available for sharing and analyzing. Preventing the disclosure of private information is a major research hot-spot. Specifically, protecting private data that is included in multi-source aggregated data is a challenge.

Taking the Internet of Vehicles (IoV) scenario as an example, hardware and software systems can detect and gather data regarding individual trajectories and environmental conditions. IoV places particular emphasis on fostering information exchanges among different entities. Vehicles can establish extensive communication using various wireless communication technologies. Figure 1 illustrates the data interaction model. Many vehicle devices are used in intelligent transportation, autonomous logistics, and smart cities [1]. However, there are several privacy challenges. For example, IoV services leverage personal data, such as location, behavioral patterns, and videos, which may inadvertently share information with third parties. Published data may leak private information about users. Addressing the publication of information related to drivers, passengers, and vehicles in IoV services has emerged as a crucial research focus, with the objective of preventing the release of sensitive information during the data publication period. Simultaneously, compliance with the General Data Protection Regulation (GDPR) [2] in the EU mandates the legal requirements.

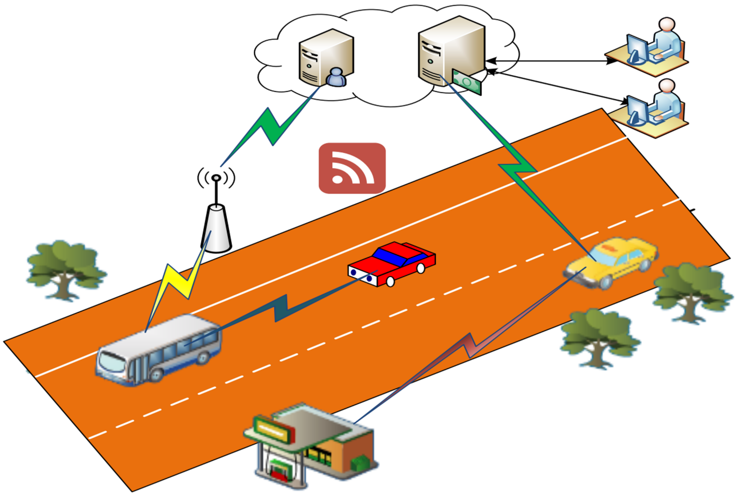


Figure 1. Internet of Vehicles data interaction model.

The *k*-anonymity technique is used to address privacy threats. Privacy-preserving data publishing (PPDP), introduced by Fung et al. in 2010 [3], has gained considerable attention in the research community and has emerged as a pivotal aspect of data mining and information sharing. Numerous studies have been devoted to anonymized data; they often consider a data publisher's table comprising explicit identifiers (EIs), quasi-identifiers (QIs), sensitive attributes (SAs), and non-sensitive attributes (NSAs). EIs refer to attributes that explicitly identify record owners. QIs are attributes that may be used to identify an individual. An instance is {gender, ZIP code, birth} that can be potentially linked to other tables for identifying the record owner. SAs are sensitive attribute such as illness and salary. NSAs are the remaining attributes [4]. Additionally, these four attribute sets are disjointed. EIs are directly linked to sensitive information; therefore, for privacy, they are typically removed before data release. However, even with the removal of EIs, record owners can potentially be re-identified, meaning their QIs can be linked to an external table [5]. To resist this attack, data publishers release an anonymized table that includes QI', SAs, and NSAs, with QI' representing the results of anonymizing QIs. The concept of *k*-anonymity is a famous method to ensure privacy protection [6] by severing the connection between specific records and their origin. This model guarantees that each record of an anonymized dataset remains indistinguishable from at least $\mathrm{k}-1$ other records [7].

Anonymization algorithms, as discussed [8,9] in various studies such as [10–13], primarily focus on optimizing efficiency and utility. However, these algorithms often overlook the impact of effectiveness and efficiency. Another category is scenario-based anonymization. Shi et al. [14] proposed "quasi-sensitive attributes" that consider that the new condition of QIs and SAs are equivalent. Terrovitis et al. [15] argued that some attributes are both quasi-identifiers (QIs) and sensitive attributes (SAs). Sei et al. [16] introduced the notion of "sensitive quasi-identifiers" for *l*-diversity and t-closeness models. These scenario-based approaches involve defining new attribute types to address various privacy objectives. Methods that address location privacy preservation include enforcing anonymized data identities [17], location obfuscation [18], space transformation [19], and spatial anonymization [20]. Bamba et al. [21] presented the PrivacyGrid framework. It anonymizes location-based queries and gives effective algorithms. Pan et al. [22] proposed ICliqueCloak for location k-anonymity.

However, protecting private data from multi-source aggregated data is complex. On the one hand, there are different types of private information, such as location, identity, and privacy information. In addition, data from many sensors are included, such as global positioning systems, radar, and cameras. Hence, many aggregated datasets have been generated from multi-source data. When information is aggregated, it is easy to leak the linked relationships. For example, an attacker can access the database of a vehicle administration office. Thus, he knows a vehicle's registration date, license plate, and color.

By observing the values in an anonymized aggregation dataset, an attacker can obtain accurate QI values from another data source. This creates new privacy disclosures in multi-source scenarios. However, current techniques must address this problem. After *k*-anonymity, some anonymized QIs of the data remain unchanged, as *k*-anonymity only entails the records in the QI group. This privacy disclosure differs from identity disclosure and attribute disclosure because it leaks more information about the attack objective, which is harmful. Considering the aggregation of datasets, it may differ from general privacy-preserving data publishing or location privacy.

Therefore, we exploit this new attack and propose countermeasures. The main contributions of this study are as follows.

- We propose a multi-source linkability attack by analyzing the problem of the aggregated dataset in IoV.
- We proposed multi-source (*k*,*d*)-anonymity and multi-source (*k*,*l*,*d*)-diversity to protect privacy disclosure in IoV. The former prevents the acquisition of accurate quasi-identifier values when the attacker possesses background knowledge. The latter has a similar privacy capability and can protect sensitive attributes. In addition, we provide heuristically efficient algorithms.
- We experimentally evaluated our algorithms using real datasets. The experimental results demonstrate that our algorithms perform well regarding privacy disclosure, information loss, and efficiency.

In Section 2, we describe related works, including the k-anonymity model and data utility of anonymity. Section 3 provides an introduction to the aggregated dataset and new attacks, elucidating the motivation behind our research. Our novel algorithms are presented in Section 4, and Section 5 showcases the results with regard to privacy, utility, and efficiency. Finally, Sections 6 and 7 discuss and conclude the study.

2. Related Works

2.1. k-Anonymity

Sweeney proposed the k-anonymity model [7,23]. This technique gained widespread popularity within the academic community. Another notable privacy model, *l*-diversity, was introduced by Machanavajjhala et al. in 2007 [24]; *l*-diversity ensures that each equivalence class contains at least *l* "well-represented" sensitive attribute values. To illustrate, consider Table 1, which is an original table detailing patient records, with *Patient Name* as an EI, *Patient Age*, *Patient Sex*, and *Patient ZIP Code* as QIs, and *Disease* as an SA. Table 2 lists the results achieved through 3-anonymity and 3-diversity. Even if data analysts possess Elle's QI values, discerning his specific records from the first three becomes challenging. Furthermore, the accurate identification of a sensitive disease is difficult because of the presence of three different values in each group. Over the years, various *k*-anonymity algorithms have been devised by employing techniques such as generalization, suppression, clustering, and micro-aggregation.

Table 1. Example of patient information.

Patient Name	Patient Age	Patient Sex	Patient ZIP Code	Patient Disease
Cart	24	Male	17227	Flu
Bob	23	Male	17672	Hepatitis
Gaul	24	Female	17537	HIV
Elle	48	Female	19240	Hangnail
Alice	51	Male	18824	Bronchitis
Helen	46	Female	18824	Flu

Table 2. Example of 3-anonymity 3-diversity patient.

Patient Age	Patient Sex (* Denotes the Value is Suppressed)	Patient Address	Patient Disease
[23–24]	*	[17226–17672]	HIV
[23–24]	*	[17226–17672]	Hepatitis
[23–24]	*	[17226–17672]	Flu
[46–51]	*	[18824–19240]	Flu
[46–51]	*	[18824–19240]	Hangnail
[46–51]	*	[18824–19240]	Bronchitis

LeFevre et al. [10] developed an efficient full-domain *k*-anonymity algorithm. It executes a breadth-first search strategy and pruning strategy to find the optimal anonymization result. Mondrian [11] proposed a partitioning algorithm that recursively partitions the domain values of QI to obtain a generalized range. Liang et al. [12] proposed optimized *k*-anonymity. It mathematically formulates the *k*-anonymity optimization problem and identifies the equivalence class with minimal information loss using an optimization solver. Shi et al. [14] introduced the concept of “quasi-sensitive attributes”. These attributes are not inherently sensitive. However, they may be sensitive when linked to an external table. Therefore, they indirectly disclose sensitive information. Terrovitis et al. [15] proposed a separation-based algorithm that addresses the condition that some attributes serve as both QIs and sensitive attributes. Sei et al. [16] introduced the sensitive quasi-identifier concept for *l*-diversity and t-closeness models. Jayapradha et al. [25] proposed heap bucketization anonymity (HBA) to protect multiple sensitive attributes while keeping balance between privacy and utility.

Onesimu et al. [26] designed a clustering-based anonymity model. It protects data for healthcare services systems. Onesimu et al. [27] proposed an attribute-focused privacy-preserving data publishing scheme. It has two different anonymity methods. The first is a fixed-interval approach, which works for numerical attributes. The last is *l*-diverse slicing for sensitive attributes. Yao et al. [28] proposed a utility-aware (α, β) privacy model, MSAAC, that can balance data privacy and utility by setting privacy parameters. Parameshwarappa et al. [29] presented a multi-level clustering-based approach to enforce an *l*-diversity model by using a non-metric weighted distance measure. Based on a clustering technique, Srijayanthi et al. [30] proposed an anonymization privacy-preserving model along with feature selection. It designs a preserved anonymization algorithm to reduce the dimensionality of the data to accelerate the action of generating clusters. Karuna and Sumalatha [31] proposed a solution that identifies the optimal seed values for aggregating similar records that can be anonymized uniformly to minimize data loss. It presents a methodical strategy for selecting seeds to cluster records by applying an adaptive k-anonymity algorithm. Guo et al. [32] proposed an entropy-based k-anonymity model. It can address static and long-term data. Prabha and Saraswathi [33] combine k-anonymity and a Laplace differential and name their technique (K, L) anonymity. It can prevent linkage attacks.

For better privacy protection of LBS information, Ma et al. [34] take into account the camouflage range and place type. Kang et al. [35] introduced MoveWithMe. It is built into a mobile app and can generate decoy queries. This app ensures that the reported movements are semantically different from the real trace. Cheng et al. [36] proposed a privacy-level allocation method that disturbs the location points before publishing in correspondence to different privacy budgets.

2.2. Anonymity Utility

Researchers designed many metrics for data utility of anonymized data. Notable measures found in the literature include the *discernibility metric* (*DM*) [37], which computes the sum of the squares of the cardinality of equivalence classes; the *classification metric*

(*CM*) [38], which involves class labels for tuple classification; the *normalized certainty penalty* (*NCP*) [39], defined as the sum of the ranges of quasi-identifiers in each equivalence class; and the *global certainty penalty* (*GCP*) [40], which normalizes the sum of the *NCP* across all equivalence classes. Notably, the *CM* is particularly suitable when anonymized data are used for decision making. These metrics effectively reflect the cardinality and domain extent of each equivalence class.

3. Models

In this section, the motivation and assumptions for the proposed attack are introduced, and a privacy model is proposed.

3.1. Notations

Consider a dataset, denoted as T, with schema $T(A_1, ..., A_v)$, where P represents the primary key, and $A_1, A_2, ..., A_v)$ are attributes, where v denotes the number of attributes. Each record in dataset T is represented as r, and the total number of tuples is denoted as n. The aggregated and anonymized datasets are denoted as AT and AT', respectively. The notation used in this study is listed in Table 3.

Table 3. Symbols and descriptions.

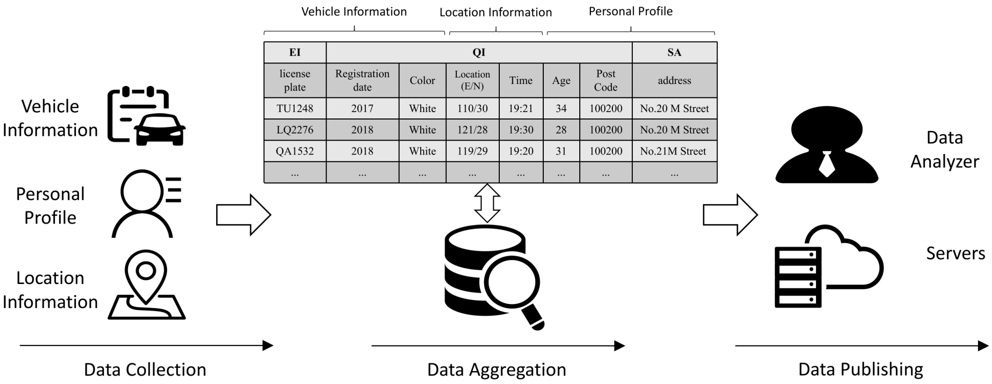
Symbol	Description
T	Original Dataset
AT	Aggregated Data
AT'	Anonymized Aggregation Data
k	K-Anonymity Parameter
r	Record
A	Attribute
v	Number of Attributes
n	Number of Records
NCP, GCP	Information Loss Metric Function

3.2. Motivation

The aggregated datasets formed in the IoV scenario can result in a new privacy problem. We begin by outlining the representation of aggregated datasets. Within IoV scenarios, data play a crucial role in supporting users through features such as automatic driving, navigation, and intelligent transportation systems. In the utilization of data, IoV services may incorporate personal information (e.g., location, behavioral patterns, and personal profiles), which could inadvertently include details about third parties, such as private properties. Additionally, IoV encompasses a variety of sensors, such as global positioning systems, radar, and cameras. Consequently, numerous aggregated datasets have been generated through the integration of data from multiple sources.

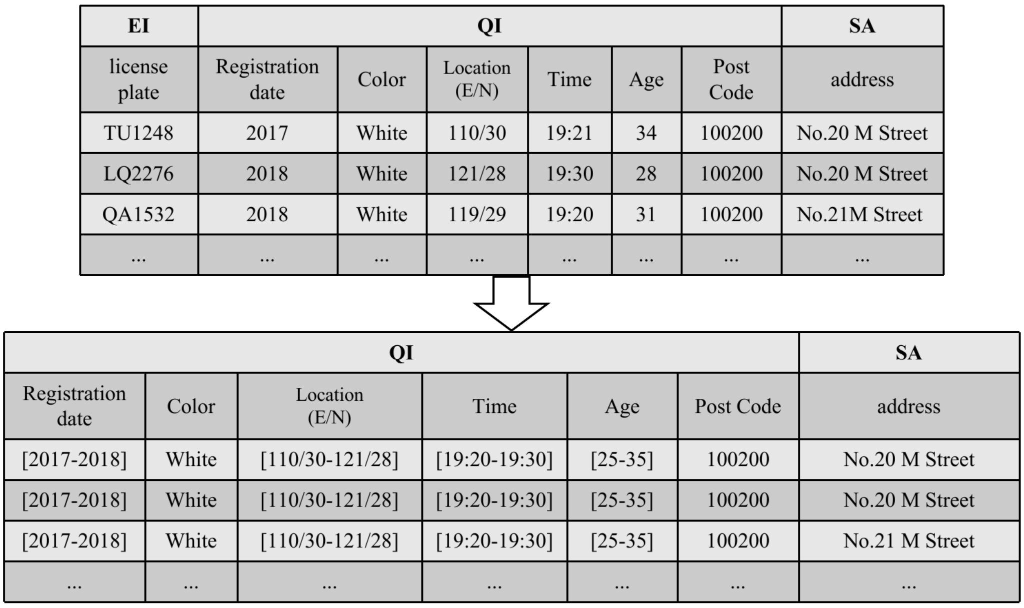
First, we describe the creation of aggregated datasets. Typically, data obtained from various devices can be utilized directly or can undergo aggregation before use. The former scenario does not lead to the formation of aggregated datasets, whereas the latter involves concatenating pertinent information about individuals, potentially leading to the exposure of sensitive details. We delve into the most straightforward case in the aggregated scenario, for which the data are from distinct sensing devices or associated databases. Figure 2 illustrates the process of publishing aggregated datasets, which can be categorized into three phases. During the data collection stage, data are gathered from different sensing devices or related databases. In the data aggregation stage, vehicle information, personal profiles, and location information are combined for publication, with each record containing information about an individual. Finally, the aggregated dataset may be released to a data analyzer or online servers for purposes such as data mining and decision making. A common practice is to anonymize private information during this stage to obfuscate sensitive information regarding the identities of individuals.

Figure 3 illustrates an aggregated dataset in the data aggregation phase; the aggregated dataset encompasses attributes such as license plates, registration dates, and colors. Here, the license plate is an EI, the address is an SA, and the others are QIs. For anonymization, the EIs are removed to prevent them from providing identification information that could reveal the unique vehicle owner. The presence of EIs is a potential source of identity disclosures. Similarly, QIs such as registration date and color, which can individually identify a unique vehicle, need to be transformed into anonymous results. For instance, values such as {2017, White, 110/30, ...} may be transformed into {[2017–2018, White, [110/30, 112/28], ...}. Figure 3 provides an example of 2-anonymity and 2-diversity for the original and anonymized datasets, respectively. In addition, a set of records sharing the same QI values constitutes an equivalence class. Failure to properly anonymize EIs or QIs may result in leakage of sensitive values. For instance, if the sensitive attribute is the address, and two records within the same equivalence class have the same anonymized QI values (e.g., No. 20 M Street), this could lead to attribute disclosure.



EI	QI						SA
license plate	Registration date	Color	Location (E/N)	Time	Age	Post Code	address
TU1248	2017	White	110/30	19:21	34	100200	No.20 M Street
LQ2276	2018	White	121/28	19:30	28	100200	No.20 M Street
QA1532	2018	White	119/29	19:20	31	100200	No.21M Street
...	...	...	...	...	...	...	...

Figure 2. Example of publishing aggregated datasets.



EI	QI						SA
license plate	Registration date	Color	Location (E/N)	Time	Age	Post Code	address
TU1248	2017	White	110/30	19:21	34	100200	No.20 M Street
LQ2276	2018	White	121/28	19:30	28	100200	No.20 M Street
QA1532	2018	White	119/29	19:20	31	100200	No.21M Street
...	...	...	...	...	...	...	...

QI						SA
Registration date	Color	Location (E/N)	Time	Age	Post Code	address
[2017-2018]	White	[110/30-121/28]	[19:20-19:30]	[25-35]	100200	No.20 M Street
[2017-2018]	White	[110/30-121/28]	[19:20-19:30]	[25-35]	100200	No.20 M Street
[2017-2018]	White	[110/30-121/28]	[19:20-19:30]	[25-35]	100200	No.21 M Street
...	...	...	...	...	...	...

Figure 3. Example of anonymized aggregation dataset.

Definition 1 (**Quasi-Identifier Attribute**). *A quasi-identifier set QI is a minimal set of attributes in original dataset T that can be joined with external information to re-identify individual record r.*

Definition 2 (**Equivalence Class**). *A set of records that contains all the same QI values, constituting an equivalence class.*

In addition to identity disclosures and attribute disclosures [41], a new form of privacy disclosure emerges; it is specifically based on linkability [42]. This form of disclosure may exist in an aggregated dataset. Linkability, in this context, refers to a data analyzer's ability to successfully discern whether two items of interest (IOIs) are linked, allowing the data analyzer to gain new information through the linkage. The definition of linkability is rooted in the literature by Pfitzmann and Hansen [43]. Linkability can potentially lead to inference. When a data analyzer links two IOIs, it may be possible to infer the actual identity from their connection.

In the IoV scenario, when data are multi-source, like related datasets or different sensors, the data from another source are valuable if the attacker owns all source data. For example, an attacker can access the database of the vehicle administration office. Thus, he knows a vehicle's registration date, license plate, and color. By observing the values in the anonymized aggregation dataset, the attacker obtains an accurate postcode value of 100200 in Figure 3, which is from another data source. This creates new privacy disclosures in multi-source scenarios. However, current techniques do not address this issue. This situation infers sensitive information, such as the postcode of the vehicle owner. K-anonymity may generalize only a few QI attributes. K-anonymity entails only the number of records in the QI group. Hence, some values remain unchanged. This privacy disclosure differs from identity and attribute disclosure because it leaks more information about the attack objective, which is harmful.

We propose a new multi-source linkability attack based on the properties of the aggregated dataset that we described.

Definition 3 (**Multi-source Linkability Attack**)**.** *A multi-source linkability attack occurs when QI data from a source are linked with some extra knowledge. These data are then linked to other QIs. Attackers can obtain accurate information from other QIs.*

The core problem is that an attacker can identify an individual based on a data source. The inference is limited to the same record; thus, the extra data belong to the same person. Any extra single piece of knowledge can expand the information about the objective.

3.3. Proposed Privacy Models

This section describes the proposed threat model and privacy model. The latter formulates and protects the proposed privacy problem in IoV.

3.3.1. Threat Model

We assume that the attacker obtains the values of the QI from one or more data sources and identifies a specific person in the original table T. Based on this knowledge, the adversary may identify a person, obtain more information from other data sources, and access sensitive values by joining knowledge and the anonymized dataset T^* to QI so that the record may be accurately linked to a specific person's sensitive attribute values. We consider the problem of generating an anonymized aggregation dataset. Therefore, the inference of an anonymized dataset should be avoided.

3.3.2. Multi-Source Anonymity

In this subsection, we describe the new privacy countermeasure. Domains are sets of terms wherein each term is an instance of a concept. Figure 4 showcases the original and generalized values for the age attribute in the taxonomy tree. As shown in the figure, the domain value of the age attribute is {0, 1, ..., 78, 79}.

A multi-source linkability attack can occur with a higher probability when the data analyzer obtains accurate values of QIs from a source. Therefore, we define a new privacy model: multi-source (k,d)-anonymity for multi-source linkability attacks.

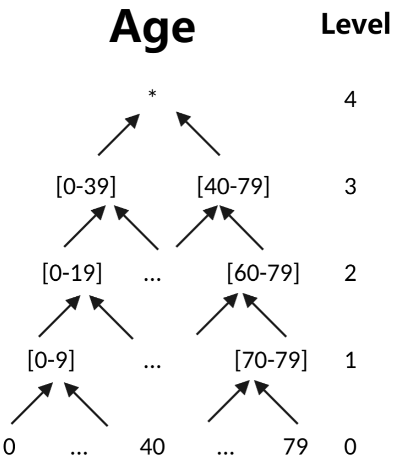


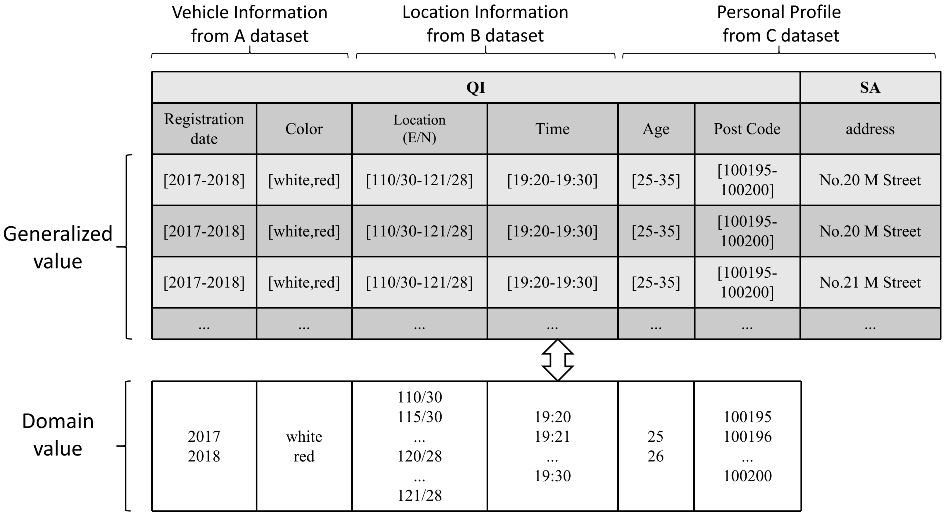
Figure 4. Examples of original and generalized values (* denotes the value is suppressed).

Definition 4 (**Multi-Source (*k*,*d*)-anonymity**)**.** *A dataset satisfies multi-source (k,d)-anonymity if each record is indistinguishable from at least k − 1 other records within the dataset and each generalized value contains at least d domain values for QI attributes. The probability of obtaining the exact values of all QIs is less than 1/d.*

In addition, we consider sensitive attributes. Multi-source (*k*,*l*,*d*)-diversity is proposed.

Definition 5 (**Multi-Source (*k*,*l*,*d*)-diversity**)**.** *A dataset satisfies multi-source (l,d)-anonymity if it satisfies multi-source (k,d)-anonymity and each indistinguishable group has at least l different values for sensitive attributes.*

Example 1. *We continue using the vehicle dataset mentioned earlier for illustration purposes. Figure 5 depicts an example dataset adhering to multi-source (3,2)-anonymity. The upper subfigure illustrates the generalized dataset meeting the (3,2)-anonymity criteria. The lower subfigure displays the domain values encompassed by the generalized values. In Figure 5, if the attacker knows that the age is 27 and the postcode is 1000195, there are at least two potential values for the color and registration date that they could infer: specifically, [2017–2018] and [white, red]. The likelihood of the data analyzer obtaining a precise combination of color and registration date is 1/2, mirroring the probability of accurately inferring from the attribute values. Consequently, attackers cannot obtain accurate values.*


Vehicle Information from A dataset
Location Information from B dataset
Personal Profile from C dataset
Generalized value
Domain value

QI						SA
Registration date	Color	Location (E/N)	Time	Age	Post Code	address
[2017-2018]	[white,red]	[110/30-121/28]	[19:20-19:30]	[25-35]	[100195-100200]	No.20 M Street
[2017-2018]	[white,red]	[110/30-121/28]	[19:20-19:30]	[25-35]	[100195-100200]	No.20 M Street
[2017-2018]	[white,red]	[110/30-121/28]	[19:20-19:30]	[25-35]	[100195-100200]	No.21 M Street
...	...	...	...	...	...	...

2017 2018	white red	110/30 115/30 ... 120/28 ... 121/28	19:20 19:21 ... 19:30	25 26	100195 100196 ... 100200

Figure 5. Example of multi-source (3,2)-anonymity.

4. Algorithms

In addition, we provide a multi-source linkability attack Algorithm 1. This can verify the privacy disclosure. It inputs the anonymized aggregated dataset AT' and the attribute set QI_s. QI_s are the attacker's data attributes. Attackers can identify individuals from QI_s and can thereby obtain accurate information by linking other QIs. The output is set as $InferenceSet$. It contains the accuracy value that the attackers obtain from other QI attributes. Lines 1–5 define parameters and variables and return anonymized equivalence class ECs. The loop of ECs checks each value of the other QI attributes. If it is the original value, it is added to $InferenceSet$. Finally, this algorithm returns all original values from the other QI attributes. The overhead of this algorithm is $O(|ECs| * m)$, where $|ECs|$ is the number of equivalence classes and m is the number of QI_d; the method is efficient.

Algorithm 1 The multi-source linkability attack algorithm.

```
Input: $AT'$,$QI_s$
Output: $InferenceSet$
 1: $InferenceSet = new\ HashMap()$
 2: $define\ EI, QI, SA\ attributes$
 3: $defineQI_s, QI_d$
 4: $ECs = EquivalenceClass(AT')$
 5: $m = the\ attribute\ number\ of\ QI_d$
 6: for $i = 0$ to $|ECs|$ do
 7:     $r = AT'[i]$
 8:     $QIkey = r.QI_s$
 9:     for $j = 0$ to $m$ do
10:         if $QI_d[j]\ is\ original\ value$ then
11:             if $!InferenceSet.contian(QIkey)$ then
12:                 $valueList = []$
13:                 $valueList.set(j, QI_d[j])$
14:                 $InferenceSet.put(QIkey, valueList)$
15:             else
16:                 $valueList = InferenceSet.get(QIkey)$
17:                 $valueList.set(j, QI_d[j])$
18:             end if
19:         end if
20:     end for
21: end for
22: return $InferenceSet$
```

We provide an implementation algorithm for the multi-source (*k*,*d*)-anonymity model. This algorithm can deliver good data utility and efficiency using the Hilbert curve. The Hilbert curve [40] is a well-known spatial mapping technique. It can map a point-of-space region to an integer. If two points are close in multidimensional space, they are likely to have similar Hilbert transform values. For example, Figure 6 illustrates the transformation of the data from 2-D to 1-D for attributes such as postcode and age. The dataset exhibits complete ordering with respect to the 1-D Hilbert values. For instance, the value {55,100195} is transformed to the 1-D value 1, whereas the value {62,100196} is transformed to a 1-D value of 63.

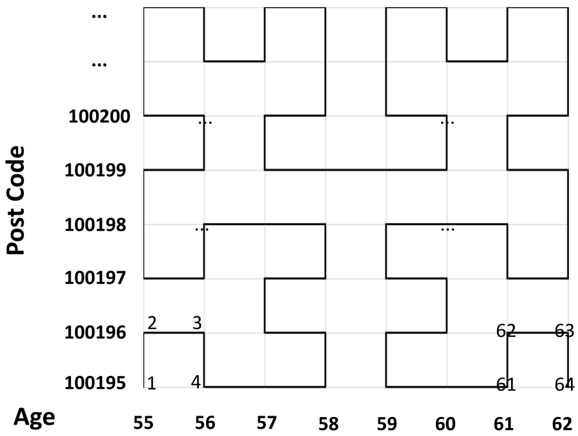


Figure 6. Example of Hilbert curve for postcode and age attributes.

In the data mapping process, each attribute value must be assigned an integer value to facilitate sorting. For numerical attributes, attribute values can be used directly owing to their inherent orderliness. As illustrated in the figure, both age and postcode can be arranged in ascending order, and each attribute value is assigned as a distinct integer. For categorical attributes, the assignment of integers is based on a taxonomy tree. For instance, considering a taxonomy tree with nodes "India" and "Japan" sharing the parent "Europe", $NCP(India, Japan) = 1/2$. On the other hand, when comparing nodes "Cambodia" and "England" with the common parent "Country", $NCP(Cambodia, England) = 1$. Therefore, distances for nodes sharing the same parent node should be smaller. Figure 7 showcases this.

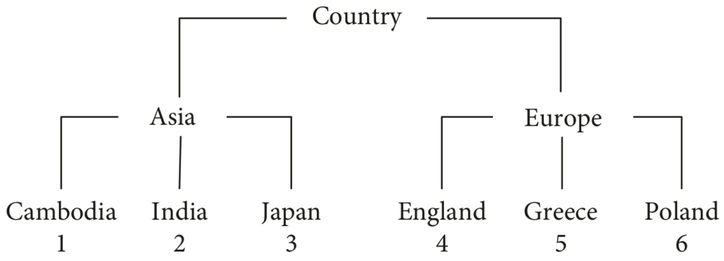


Figure 7. Taxonomy tree of country attribute.

We propose an efficient and heuristic multi-source (*k*,*d*)-anonymity algorithm. Algorithm 2 describes the primary implementation process. The algorithm inputs the original aggregated dataset AT and anonymity requirements k, t. First, it defines the EIs, QIs, and SAs. Subsequently, the EIs are deleted to satisfy the anonymization requirements. The term d_{QI} denotes the number of QIs. $Domains$ are a set containing domain values of the QIs. Lines 4–7 obtain the domain values. The $DOMAIN$ function obtains all domain values of QI_i. Based on the domain value $Domain$ and dimension d_{QI}, a Hilbert curve is plotted on AT. It implements the mapping from d_{QI}-D to 1-D by Hilbert transform. It sort the records with d_{QI}-D attributes in the dataset AT. Then, the loop is entered until $AT = \emptyset$. In the loop, if the number of records is less than k or the distinct domain value of a QI is less than d, the loop is stopped. Lines 15–18 iterate through the sorted data and obtain record r from dataset AT. The loop determine whether the set $Records$ satisfies the condition of anonymity. If this condition is satisfied, the group is anonymized by generalizing the QI values. An anonymized equivalence class is generated. In the last step, anonymized values are incorporated into the set AT'. The ultimate output is AT', which encompasses all anonymized records that adhere to anonymity requirements k and d. The computational overhead of this algorithm is $O(d)$, which ensures its efficiency. The I/O cost is linear.

Algorithm 3 shows the multi-source (*k*,*l*,*d*)-anonymity algorithm. The algorithm inputs the original aggregated dataset AT and anonymity requirements k, l, and t. First, it defines the EIs, QIs, and SAs. Subsequently, the EIs are deleted to satisfy the requirements of anonymization. The term d_{QI} denotes the number of QIs. *Domains* is a set containing domain values of the QIs. Lines 5–7 obtain the domain values. The $DOMAIN$ function obtains all the domain values of QI_i. $SADomain$ is the domain value of the SA. *Curve* implements the mapping from d_{QI}-D to 1-D using the Hilbert transform. The records are sorted according to their SA values and assigned to m domains based on the sensitive attribute value of the SA domain. Let F be the set of the first records in each bucket of H. Then, the loop is entered until $AT = \emptyset$. In the loop, if the number of records is less than k or the distinct domain value of a QI is less than d, the loop stops. Lines 18–21 iterate through the sorted data and take the record r from the dataset AT. The loops then determines whether the set *Records* satisfies the condition of (k,d)-anonymity. In addition, if the distinct SA values of this group are less than l, another loop continues. Lines 22–35 get records with different SA values and the minimum *curvePoint* in F. After adding records with different SA values, if the condition of this group is satisfied, the values are anonymized by generalizing the QIs. The anonymized equivalence class is created, and, ultimately, the output AT' includes all anonymized records that meet the anonymity requirements k, d, and l. With a computational overhead of $O(d)$, the method proves to be efficient. Given that the input dataset AT is ordered, the proposed approach requires only a single pass through the data. The I/O cost is linear.

Algorithm 2 The multi-source (k,d)-anonymity algorithm.

```
Input: $AT, k, d$
Output: $AT'$
1:  $define\ EI, QI, SA\ attributes$
2:  $delete\ EI\ attributes\ from\ AT$
3:  $d_{QI} = number\ of\ QI\ attribute$
4:  $Domains = \{\}$
5:  for $i = 0$ to $d_{QI}$ do
6:      $domValues = DOMAIN(QI_i)$
7:      $Domain_i.addAll(domValues)$
8:  end for
9:  $Curve \leftarrow Hilbert(Domains, d_{QI})$
10: while $|AT| \neq \emptyset$ do
11:     if $|AT| <= k\ or\ distinct\ domain\ value\ of\ QI < d$ then
12:         $Break$
13:     end if
14:     $Records = \{\}$
15:     while $|Records| < k\ and\ distinct\ domain\ value\ of\ QI < d$ do
16:         $r = Curve.nextPoint()$
17:         $Records.add(r)$
18:     end while
19:     $anonymized(Records)$
20:     $AT'.addAll(Records)$
21: end while
22: return $AT'$
```

Algorithm 3 The multi-source (*k*,*l*,*d*)-diversity algorithm.

```
Input: $AT, k, l, d$
Output: $AT'$
$define\ EI, QI, SA\ attributes$
$delete\ EI\ attributes\ from\ AT$
$d_{QI} = number\ of\ QI\ attribute$
$Domains = \{\}$
for $i = 0$ to $d_{QI}$ do
    $domValues = DOMAIN(QI_i)$
    $Domain_i.addAll(domValues)$
end for
$SADomain = DOMAIN(SA)$
$Curve \leftarrow Hilbert(Domains, d_{QI})$
$H[i]$ = Split records of Curve into m buckets based on SADomain value
F = set of first record in each bucket
while $|AT| \neq \emptyset$ do
    if $|AT| <= k\ or\ distinct\ domain\ value\ of\ QI < d$ then
        $Break$
    end if
    $Records = \{\}$
    while $|Records| < k\ and\ distinct\ domain\ value\ of\ QI < d$ do
        $r = Curve.nextPoint()$
        $Records.add(r)$
    end while
    while $distinct\ value\ of\ SA < l$ do
        $SAvalues = getAllSAValue(Records)$
        $index = -1, 1DNumber = MAXVALUE$
        for $i = 0$ to $m$ do
            if $!SAvalues.contain(SADomain[i])$ then
                $curvePoint = curvePoint\ of\ F[i]$
                if $curvePoint < 1DNumber$ then
                    $index = i, 1DNumber = 1D$
                end if
            end if
        end for
        $Records.add(F[i])$
        $updateSet(F)$
    end while
    $anonymized(Records)$
    $AT'.addAll(Records)$
end while
return $AT'$
```

5. Experimental Evaluation

We conduct an experimental evaluation to assess the algorithms' performance regarding privacy, data utility, and efficiency. Section 5.2 presents the evaluation of privacy disclosure in the multi-source dataset. Section 5.3 provides experimental results showing the data utility achieved by our algorithms. Finally, in Section 5.4, we present experiments focused on assessing the efficiency of our algorithms.

5.1. Experiment Description

Dataset: We evaluate our algorithms using publicly available datasets. The Adult dataset from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/2/adult (accessed on 10 September 2023)) and the Census dataset from the US full 1990 census (https://archive.ics.uci.edu/dataset/116/us+census+data+1990 (accessed on 10 September 2023)) are selected. These datasets are the de facto standards for the evaluation of anonymization. The Census dataset contains a one-percent sample of the Public Use Microdata Samples (PUMS) person records. The Adult dataset consists of 32,562 records with 14 attributes, of which three are numerical and the rest are categorical. The Census dataset consists of 2,458,285 records. All attributes are categorical. Six attributes from the Adult and Census datasets were used in our experiment. For the Adult dataset, we consider {age, workclass, marital} as attribute set QI_1 and {occupation, relationship, native-country} as attribute set QI_2. For the Census dataset, we consider {dAge, dAncstry1, dAncstry2} as attribute set QI_1 and {iClass, dDepart, dHispanic } as attribute set QI_2. The experiments simulate the condition of aggregated datasets from two different sources.

Experimental environment and algorithms: The experiments are carried out on a machine featuring a 3.0 GHz Intel(R) Core(TM) i5 processor with 12 GB RAM. The operating system used is Ubuntu 22.04, and the implementation is developed and executed on an IntelliJ IDEA 2023. Java is used as the programming language, and the JDK version is 15. In addition, we implement the well-known Top-down [44], Mondrian [11], Incognito [10], and Flash [45] anonymity as our compared algorithms for privacy disclosure experiments. The experiments are repeated ten times, and the average result for each trial is calculated. The default parameters {*k*,*d*,*l*} are 20, 2, 3 and 200, 2, 3 on the Adult and Census datasets, respectively.

Experimental objectives: We have three main objectives for the following experiments: **Privacy Disclosure Evaluation—**this evaluation tests the effects of a multi-source linkability attack using Algorithm 2. This verifies the effectiveness of the proposed method. **Data utility—**the experiments focus on information loss of the proposed anonymity algorithm under different conditions. **Efficiency—**experiments are conducted to evaluate the required overhead with different numbers of records.

5.2. Privacy Disclosure

In this subsection, we present the experimental privacy disclosure results for the aggregated datasets from different sources. QI_1 and QI_2 are two QI attribute sets from two source. Assume that the attacker can acquire background knowledge of the original QI_1 or QI_2. The multi-source linkability attack algorithm (Algorithm 2) can get accurate original values of another QI set. The experiment results are obtained via executing Algorithm 2. We test privacy disclosure on the Adult dataset and the Census dataset.

Tables 4 and 5 present the privacy disclosures of k-anonymity and l-diversity, respectively, on the Adult dataset. Our techniques are multi-source (*k*,*d*)-anonymity and multi-source (*k*,*l*,*d*)-diversity. As described above, for the Adult dataset, the {a1, a2, a3} attributes are {age, workclass, marital}, and for {a4, a5, a6}, the attributes are {occupation, relationship, native-country}. The values in {a1, ..., a6} represent the number of distinct original values in an equivalence class. For example, if an equivalence class has {[40–59], Private,...} in {a1, ..., a6}, a2 has a distinct original value. However, a1 has a generalized value that cannot leak accurate age information. The rows of {a1, ..., a6} store distinct original values. Top-down has the most equivalence classes, and its privacy disclosure is significant. The sum of QI_1 and QI_2 is 6085. Mondrian has fewer privacy disclosures. For the Incognito and Flash algorithms, the results are interesting and different. These two algorithms generate only 10–200 equivalence classes. Hence, the number of distinct original values is small. Incognito has zero, in that all values in an equivalence class are range values instead of accurate values. Our algorithm performs well, with zero distinct original values. Table 5 presents the l-diversity algorithm. Incognito and Flash can execute l-diversity. These results are similar to those shown in Table 5.

Table 4. Privacy disclosures of k-anonymity on adult dataset.

Algorithm	QI_1 Set			QI_2 Set			Number of Equivalence Classes	Privacy Leakage of QI_1 Set	Privacy leakage of QI_2 Set	Privacy Leakage
	a1	**a2**	**a3**	**a4**	**a5**	**a6**				
Top-down [44]	0	1206	1364	1098	1242	1175	1364	2570	3515	6085
Mondrian [11]	2	696	646	122	622	538	987	1344	1282	2626
Incognito [10]	0	0	0	0	0	0	15	0	0	0
Flash [45]	1	18	184	18	184	6	184	203	208	411
Our	0	0	0	0	0	0	388	0	0	0

Table 5. Privacy disclosures of l-diversity on Adult dataset.

Algorithm	QI_1 Set			QI_2 Set			Number of Equivalence Classes	Privacy Leakage of QI_1 Set	Privacy Leakage of QI_2 Set	Privacy Leakage
	a1	**a2**	**a3**	**a4**	**a5**	**a6**				
Incognito [10]	0	0	0	0	0	0	15	0	0	0
Flash [45]	1	18	184	18	184	6	184	203	208	411
Our	0	0	0	0	0	0	389	0	0	0

Tables 6 and 7 present the privacy disclosures of k-anonymity and l-diversity on the Census dataset. As the number of records increases, the number of equivalence classes and privacy leaks increase. These results are similar.

Table 6. Privacy disclosures of k-anonymity on Census dataset.

Algorithm	QI_1 Set			QI_2 Set			Number of Equivalence Classes	Privacy Leakage of QI_1 Set	Privacy Leakage of QI_2 Set	Privacy Leakage
	a1	**a2**	**a3**	**a4**	**a5**	**a6**				
Mondrian [11]	823	1679	2467	1534	1539	2583	2920	4969	5656	10,625
Incognito [10]	0	0	0	0	76	0	76	76	76	152
Flash [45]	1	818	818	818	818	818	818	1637	2454	4091
Our	0	0	0	0	0	0	137	0	0	0

Table 7. Privacy disclosures of l-diversity on Census dataset.

Algorithm	QI_1 Set			QI_2 Set			Number of Equivalence Classes	Privacy Leakage of QI_1 Set	Privacy Leakage of QI_2 Set	Privacy Leakage
	a1	**a2**	**a3**	**a4**	**a5**	**a6**				
Incognito [10]	0	0	0	0	76	0	76	76	76	152
Flash [45]	1	818	818	818	818	818	818	1637	2454	4091
Our	0	0	0	0	0	0	138	0	0	0

5.3. Data Utility

In this subsection, we show the results of multi-source (k,d)-anonymity and multi-source (k,l,d)-diversity in terms of data utility. We measure the information loss of the anonymized dataset using the global certainty penalty and discernibility metric. Figure 8a shows the GCP information loss metric for the Adult and Census datasets. Figure 8b shows the DM information loss metric. The parameters {k,d,l} are 20, 2, 3 and 200, 2, 3 on the adult and census datasets, respectively. The higher the GCP and DM, the greater the information loss. Multi-source (k,d)-anonymity and multi-source (k,l,d)-diversity have a similar result on two different datasets for GCP and DM. Multi-source (k,d)-anonymity outperforms multi-source (k,l,d)-diversity in that privacy constraints are easier to achieve. Figure 9a–d shows the variation in the information loss with increasing k values for the two algorithms. Figure 9a,b show the GCP metric on the Adult and Census datasets, respectively. Then, Figure 9c,d show the DM metric on the Adult and Census datasets, respectively. As shown in the figure, for all k values (10 <= k <= 20 for the Adult dataset and 100 <= k <= 200 for the Census dataset), these two algorithms have similar results in terms of GCP and DM. Meanwhile, the information loss of the two algorithms remains stable with increasing k values.

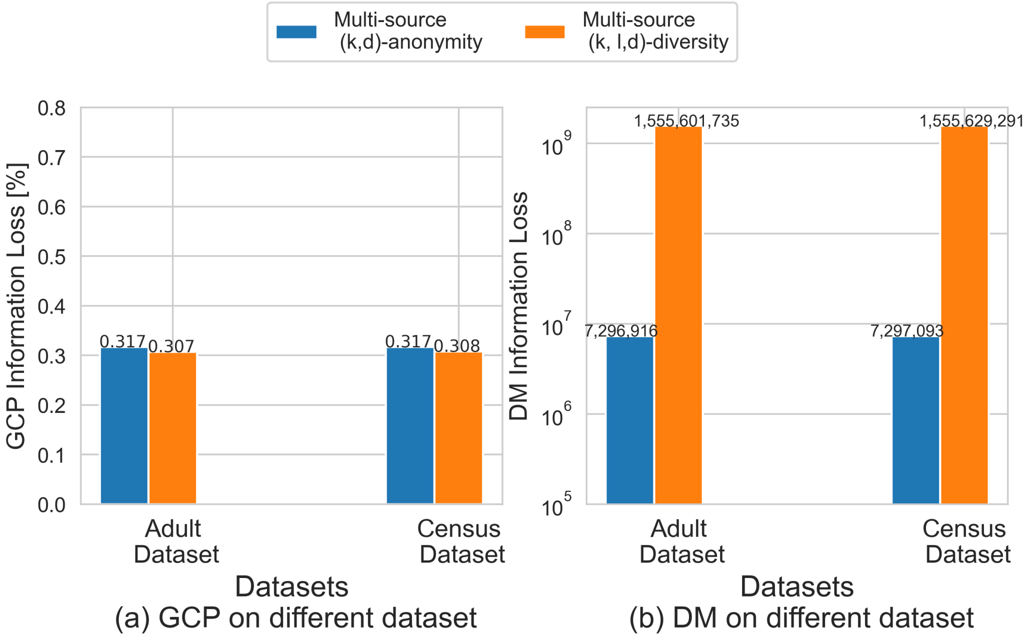


Figure 8. Information loss on Adult and Census datasets.

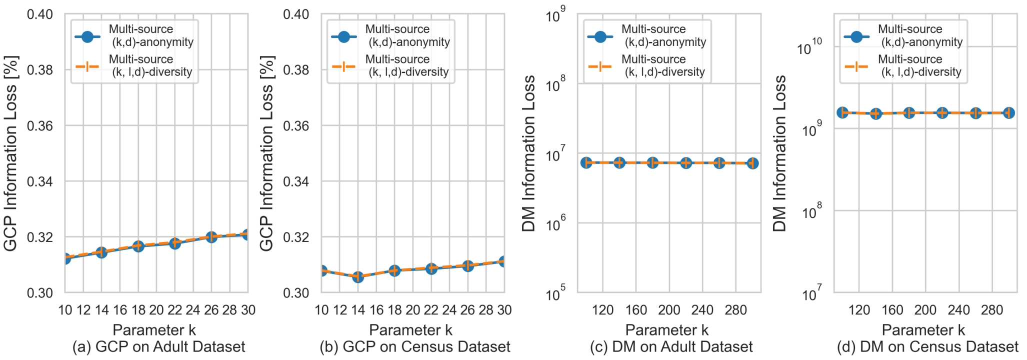


Figure 9. Information loss for different values of k.

In Figure 10a–d, we vary parameter d. That is the privacy parameter of multi-source (*k*,*d*)-anonymity and multi-source (*k*,*l*,*d*)-diversity. Figure 10a,b show the GCP metric on the Adult and Census datasets, respectively. Then, Figure 10c,d show the DM metric on the Adult and Census datasets, respectively. As the value of d increases, GCP and DM sharply increase. When d is 1, the GCP is close to 0.12–0.14 for multi-source (*k*,*d*)-anonymity and multi-source (*k*,*l*,*d*)-diversity. And when d is 5, the results are high: 0.6 and 0.8, respectively. For the DM metric, the changes were also large, ranging from 10^6 to 10^8 on the Adult dataset. In Figure 11, we compare algorithms with different values of l. Figure 11a,b show the GCP metric on the Adult and Census datasets, respectively. Then, Figure 11c,d show the DM metric on the Adult and Census datasets, respectively. These experiments can only perform multi-source (*k*,*l*,*d*)-diversity). The results show that the trend is steady for different values of l using the GCP and DM metrics.

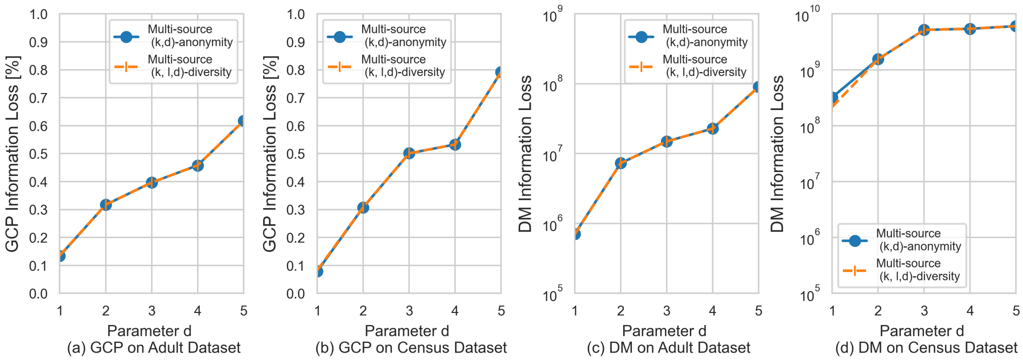


Figure 10. Information loss for different values of d.

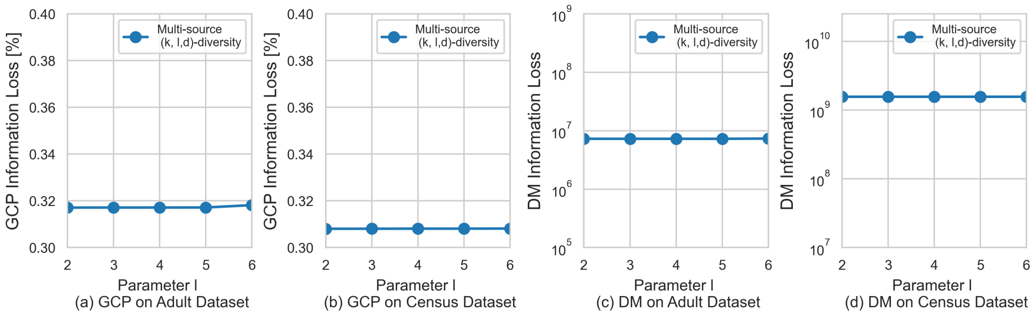


Figure 11. Information loss for different values of l.

5.4. Efficiency

Table 8 presents a comparison of the execution times in milliseconds. We can consider the Adult dataset (32,562 records) and the Census dataset (2,458,285 records) as small and large datasets, respectively. The execution time results do not include reading the original data from the disk or writing generalized data to the disk. For the Adult dataset, the execution time is approximately 0.3 s. For the Census dataset, the execution time is approximately 61–68 s. Multi-source (*k*,*l*,*d*)-diversity spends more time compared to multi-source (*k*,*d*)-anonymity. Given the superior quality of the results, the running time of our algorithms is acceptable in practice. Figure 12 shows the scalability experiment. It evaluates scalability by changing the number of QIs and increasing the size of records on the Census dataset. The different records and QIs did not significantly affect the execution time. We vary QIs and the size of the records, and the running time curve remains steady.

Table 8. Comparison of average execution times.

Dataset\Algorithm	Multi-Source (*k*,*d*)-Anonymity	Multi-Source (*k*,*l*,*d*)-Diversity
Adult	300	332
Census	61,899	67,715

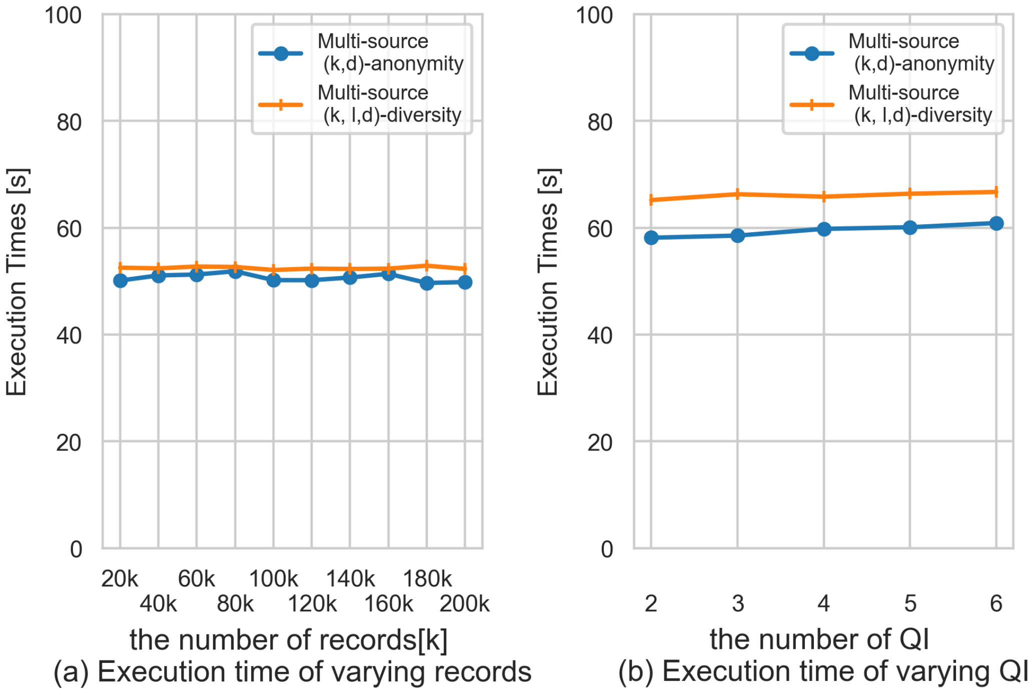


Figure 12. Scalability experiments.

6. Discussion

Anonymization and differential privacy [46] are two distinct privacy-preserving techniques in data science and analysis. Differential privacy is a mathematical framework for ensuring individual privacy when analyzing or releasing statistical data. It provides a guarantee that the inclusion or exclusion of any single individual's data will have a minimal impact on the output of an analysis. By adding carefully calibrated "noise" to query results, differential privacy ensures that no one, not even the data holder, can confidently infer whether any particular individual's data are part of the dataset. This approach allows for meaningful aggregated insights while protecting against attacks that might try to uncover sensitive information about specific individuals.

Anonymization refers to the process of removing or obfuscating personally identifiable information from datasets so that individuals cannot be recognized. This typically involves techniques like aggregation, generalization, or suppression of attributes. The goal is to create a dataset wherein the connections between specific records and their original subjects are severed.

In summary, anonymization focuses on removing direct identifiers, whereas differential privacy injects controlled randomness into data processing to limit the leakage of information. They are different tools to provide privacy protection for data.

7. Conclusions and Future Work

In this study, we explore a new attack method: a multi-source linkability attack. It occurs when QI data from a source are linked to some extra knowledge. These data are then linked to other QIs. Hence, attackers can obtain accurate information from other QIs. This paper presents a new perspective on addressing the multi-source linkability problem of the IoV scenario. We describe its implementations and cases of this new attack on a multi-source dataset.

Then, we provide two new privacy models. We extend k-anonymity to multi-source (k,d)-anonymity and multi-source (k,l,d)-diversity for two different targets. The former solves the problem of protecting privacy through multi-source QIs, while the latter simultaneously focuses on multi-source QIs and SAs. The algorithms of the above models are based on the Hilbert curve, and they are efficient on two real-world datasets. Through experiments on the real datasets, we demonstrate that our algorithms are effective for new privacy challenges. We also experimentally test privacy disclosure, data utility, and efficiency.

In the future, we will study the following problems: better support for other models and higher efficiency. Specifically, we will explore further privacy problems, because in the IoV scenario, aggregated datasets change the relations among multi-source data. Further, higher efficiency can improve the practicability of our algorithms.

Author Contributions: Methodology, X.L.; writing—original draft preparation, X.L.; writing—review and editing, W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Sadiku, M.N.; Tembely, M.; Musa, S.M. Internet of vehicles: An introduction. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2018**, *8*, 11. [CrossRef]
2. General Data Protection Regulation (GDPR). 2016. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 (accessed on 10 September 2023).
3. Fung, B.C.; Wang, K.; Fu, A.W.C.; Philip, S.Y. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2010.
4. Fung, B.C.; Wang, K.; Chen, R.; Yu, P.S. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv. (CSUR)* **2010**, *42*, 1–53. [CrossRef]
5. Sweeney, L. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 571–588. [CrossRef]
6. Samarati, P. Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **2001**, *13*, 1010–1027. [CrossRef]
7. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570. [CrossRef]
8. Oh, S.R.; Seo, Y.D.; Lee, E.; Kim, Y.G. A comprehensive survey on security and privacy for electronic health data. *Int. J. Environ. Res. Public Health* **2021**, *18*, 9668. [CrossRef] [PubMed]
9. Olatunji, I.E.; Rauch, J.; Katzensteiner, M.; Khosla, M. A review of anonymization for healthcare data. *Big Data* 2022, *online ahead of print*.
10. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Incognito: Efficient full-domain k-anonymity. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, 14–16 June 2005; pp. 49–60.
11. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 3–7 April 2006; p. 25.
12. Liang, Y.; Samavi, R. Optimization-based k-anonymity algorithms. *Comput. Secur.* **2020**, *93*, 101753. [CrossRef]
13. 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]
14. Shi, P.; Xiong, L.; Fung, B.C. Anonymizing data with quasi-sensitive attribute values. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada, 26–30 October 2010; pp. 1389–1392.
15. Terrovitis, M.; Mamoulis, N.; Liagouris, J.; Skiadopoulos, S. Privacy Preservation by Disassociation. *Proc. VLDB Endow.* **2012**, *5*, 944–955. [CrossRef]
16. Sei, Y.; Okumura, H.; Takenouchi, T.; Ohsuga, A. Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness. *IEEE Trans. Dependable Secur. Comput.* **2017**, *16*, 580–593. [CrossRef]
17. Freudiger, J.; Manshaei, M.H.; Hubaux, J.; Parkes, D.C. Non-Cooperative Location Privacy. *IEEE Trans. Dependable Secur. Comput.* **2013**, *10*, 84–98. [CrossRef]
18. Li, M.; Salinas, S.; Thapa, A.; Li, P. n-CD: A geometric approach to preserving location privacy in location-based services. In Proceedings of the IEEE INFOCOM 2013, Turin, Italy, 14–19 April 2013; pp. 3012–3020.

19. Ghinita, G.; Kalnis, P.; Khoshgozaran, A.; Shahabi, C.; Tan, K. Private queries in location based services: Anonymizers are not necessary. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, 10–12 June 2008; pp. 121–132.
20. Hoh, B.; Iwuchukwu, T.; Jacobson, Q.; Work, D.B.; Bayen, A.M.; Herring, R.; Herrera, J.C.; Gruteser, M.; Annavaram, M.; Ban, J. Enhancing Privacy and Accuracy in Probe Vehicle-Based Traffic Monitoring via Virtual Trip Lines. *IEEE Trans. Mob. Comput.* **2012**, *11*, 849–864. [CrossRef]
21. Bamba, B.; Liu, L.; Pesti, P.; Wang, T. Supporting anonymous location queries in mobile environments with privacygrid. In Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, 21–25 April 2008; pp. 237–246.
22. Pan, X.; Xu, J.; Meng, X. Protecting Location Privacy against Location-Dependent Attacks in Mobile Services. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 1506–1519. [CrossRef]
23. Samarati, P.; Sweeney, L. *Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression*; technical report; SRI International: Menlo Park, CA, USA, 1998.
24. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. *L*-diversity: Privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3. [CrossRef]
25. Jayapradha, J.; Prakash, M.; Alotaibi, Y.; Khalaf, O.I.; Alghamdi, S.A. Heap Bucketization Anonymity—An Efficient Privacy-Preserving Data Publishing Model for Multiple Sensitive Attributes. *IEEE Access* **2022**, *10*, 28773–28791. [CrossRef]
26. Onesimu, J.A.; Karthikeyan, J.; Sei, Y. An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services. *Peer-Peer Netw. Appl.* **2021**, *14*, 1629–1649. [CrossRef]
27. Onesimu, J.A.; J, K.; Eunice, J.; Pomplun, M.; Dang, H. Privacy Preserving Attribute-Focused Anonymization Scheme for Healthcare Data Publishing. *IEEE Access* **2022**, *10*, 86979–86997. [CrossRef]
28. Yao, L.; Wang, X.; Hu, H.; Wu, G. A Utility-aware Anonymization Model for Multiple Sensitive Attributes Based on Association Concealment. *IEEE Trans. Dependable Secur. Comput.* **2023**, 1–12. [CrossRef]
29. Parameshwarappa, P.; Chen, Z.; Koru, G. Anonymization of Daily Activity Data by Using l-diversity Privacy Model. *ACM Trans. Manage. Inf. Syst.* **2021**, *12*, 1–21. [CrossRef]
30. Srijayanthi, S.; Sethukarasi, T. Design of privacy preserving model based on clustering involved anonymization along with feature selection. *Comput. Secur.* **2023**, *126*, 103027. [CrossRef]
31. Arava, K.; Lingamgunta, S. Adaptive k-anonymity approach for privacy preserving in cloud. *Arab. J. Sci. Eng.* **2020**, *45*, 2425–2432. [CrossRef]
32. Guo, J.; Yang, M.; Wan, B. A Practical Privacy-Preserving Publishing Mechanism Based on Personalized k-Anonymity and Temporal Differential Privacy for Wearable IoT Applications. *Symmetry* **2021**, *13*, 1043. [CrossRef]
33. Mohana Prabha, K.; Vidhya Saraswathi, P. Suppressed K-Anonymity Multi-Factor Authentication Based Schmidt-Samoa Cryptography for privacy preserved data access in cloud computing. *Comput. Commun.* **2020**, *158*, 85–94. [CrossRef]
34. Ma, C.; Yan, Z.; Chen, C.W. SSPA-LBS: Scalable and Social-Friendly Privacy-Aware Location-Based Services. *IEEE Trans. Multim.* **2019**, *21*, 2146–2156. [CrossRef]
35. Kang, J.; Steiert, D.; Lin, D.; Fu, Y. MoveWithMe: Location Privacy Preservation for Smartphone Users. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 711–724. [CrossRef]
36. Cheng, W.; Wen, R.; Huang, H.; Miao, W.; Wang, C. OPTDP: Towards optimal personalized trajectory differential privacy for trajectory data publishing. *Neurocomputing* **2022**, *472*, 201–211. [CrossRef]
37. Bayardo, R.J.; Agrawal, R. Data privacy through optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; pp. 217–228.
38. Iyengar, V.S. Transforming data to satisfy privacy constraints. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 279–288.
39. Xu, J.; 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, Philadelphia, PA, USA, 20–23 August 2006; pp. 785–790.
40. Ghinita, G.; Karras, P.; Kalnis, P.; Mamoulis, N. Fast data anonymization with low information loss. In Proceedings of the 33rd International Conference on Very Large Data Bases, Vienna, Austria, 23–27 September 2007; pp. 758–769.
41. Prasser, F.; Bild, R.; Eicher, J.; Spengler, H.; Kuhn, K.A. Lightning: Utility-Driven Anonymization of High-Dimensional Data. *Trans. Data Priv.* **2016**, *9*, 161–185.
42. Wuyts, K.; Joosen, W. LINDDUN privacy threat modeling: A tutorial. In *CW Reports*; KU Leuven: Leuven, Belgium, 2015.
43. Pfitzmann, A.; Hansen, M. *A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management*; TU Dresden: Dresden, Germany, 2010.
44. Fung, B.C.; Wang, K.; Yu, P.S. Top-down specialization for information and privacy preservation. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; pp. 205–216.

45. Kohlmayer, F.; Prasser, F.; Eckert, C.; Kemper, A.; Kuhn, K.A. Flash: Efficient, stable and optimal k-anonymity. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; pp. 708–717.
46. Dwork, C. Differential privacy. In Proceedings of the International Colloquium on Automata, Languages, and Programming, Venice, Italy, 10–14 July 2006; pp. 1–12.

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.