A Comparative Study for Anonymizing Datasets with Multiple Sensitive Attributes and Multiple Records

Mona Mohamed Nasr[1], Hayam Mohamed Sayed[2], and Waleed Mahmoud Ead[2(✉)]

[1] Information Systems Department, Faculty of Computers and Artificial Intelligence, Helwan University, Helwan, Egypt

[2] Information Systems Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt

{hoyam.mohamed25,waleedead}@fcis.bsu.edu.eg

Abstract. Today, there are many sources of data, such as IoT devices, that produce a massive amount of data, particularly in the healthcare industry. This microdata needs to be published, and shared for medical research purposes, data analysis, mining, learning analytics tasks, and the decision-making process. But this published data contains sensitive and private information for individuals, and if this microdata is published in its original format, the privacy of individuals may be disclosed, which puts the individuals at risk, especially if an adversary has strong background knowledge about the target individual. Owning multiple records and multiple sensitive attributes (MSA) for an individual can lead to new privacy leakages or disclosure. So, the fundamental issue is how to protect the privacy of 1:M with the MSA dataset using anonymization techniques and methods, as well as how to balance utility and privacy, for this data while reducing information loss and misuse. The objective of this paper is to use different methods and different anonymization algorithms, like the 1:m-generalization algorithm and Mondrian, and compare them to show which of them maintains data privacy and high utility of analysis results at the same time. From this comparison, we found that the m-generalization algorithm and the (p, k) angelization method perform well in terms of information loss and data utility compared to the other remaining methods and algorithms.

Keywords: Privacy · Anonymization · Healthcare · Data publishing · MSA · PPDP · IOT

1 Introduction

The IOT and storage technologies' rapid development results in the collection and integration of a flood of data and types of digital information available, as the data is being generated everywhere through various sources and organizations like healthcare, Biomedical, finance, social media, and so on. This collected data needs to be published, as medical data sharing is becoming a big demand for the purpose of analysis

R. Hou et al. (Eds.): BDTA 2021/2022, LNICST 480, pp. 38–57, 2023.
https://doi.org/10.1007/978-3-031-33614-0_3

and mining tasks in order to generate and process useful patterns. These patterns help managers and researchers make decisions about studying the characteristics of diseases and discovering new drugs. Besides this usefulness, the data may contain some sensitive information, which raises a privacy concern. As a result, determining how to protect privacy and prevent privacy leakage has been a significant challenge.

To address the trade-offs between data analysis and maintaining the privacy of specific information about individuals, such as their personally identifiable attributes or explicit attributes (like name, Zip code, address, social security number, and contact number), quasi-identifier attributes (like age, gender, and), and sensitive attributes (e.g. salary and disease), privacy-preserving data mining is proposed [8].

By using electronic health records (EHRs), numerous patient data records have been used in biomedical research projects. This data needs to be secure and protected from privacy disclosure, misuse, and exploitation. For example, in the era of IoT, new wearable devices like the Apple iWatch and Google Fit can give sensitive information about individuals like their location, health condition, and financial status by recording and analyzing their daily activities.

A specific person or group of persons can then be identified using this information, either alone or in combination. Especially when the attackers have strong background knowledge that helps them make correlation attacks and infer the privacy of individuals. For instance, many researches showed that a "linking attack" could identify almost 87% (Sweeney 2000) of the US population (i.e., connecting quasi identification properties with external information like a voter list) [2, 32].

To protect individual privacy, removing personally identifying information from databases is insufficient. As a result, to resolve these privacy concerns, many PPDP (privacy-preserving data publishing) algorithms are implemented by researchers. The most popular PPDP techniques are anonymization techniques because they have less information loss and high data utility. The requirements of the data miner and analyst are therefore taken into account for anonymization in order to produce anonymized data with significantly higher utility. Because maintaining privacy while increasing utility are two contradictory demands in data sharing, or contradictory goals, this means we have to lose one to gain the other. Further, the privacy preservation method is also dependent on the type or structure of the underlying dataset. At present, most privacy-preserving models for publishing data are applied to a dataset with just one record for each individual and a single sensitive attribute [15]. However, an individual may have multiple records and multiple sensitive attributes in a dataset. For instance, a user may post many status updates or messages using the same account on social networking sites (like Facebook and Twitter). In the same dataset, a patient may have records for more than one diagnosis. In this study, we employ 1: M generalization, a technique for anonymizing 1: M datasets. There are different models for preserving privacy, like l-diversity and SLAMS. However, the L-diversity model ignores the semantic relationship between sensitive values. So, there is a chance that the data may be disclosed. So we used different methods and anonymization algorithms to show which method did well in terms of privacy and utility, such as 1: M generalization, which performs SA anonymization first, followed by QID generalization. Use the Partition [30] for sensitive attribute anonymization and for QID generalization using the Mondrian [31]. Because semantic anonymization relies on

semantic rules that lead to similarity attacks, we apply the appropriate anonymization process to such rules to avoid similarity attacks by checking each equivalence class of sensitive attributes. Other methods, such as Rating Privacy Preservation for MSA with Different Sensitivity Requirements, are made available in this manner, depending on the various sensitivity coefficients for various attributes [25]. This strategy preserves a lot of correlations between the microdata while also protecting the privacy of many sensitive variables.

(p-k) angelization for MSA, which uses the weight calculation to determine the correlation between sensitive attributes, the (p, k)-Angelization strategy decreases information loss while also protecting privacy by removing the risk of background join and non-membership attacks [15], but it takes more execution time (and has less accuracy) due to the complicated process of weight calculation.

We make use of two actual datasets. (heart disease, Informs), from uci –machine learning, (https://sites.google.com/site/informsdataminingcontest). Applying the mentioned methods to these datasets to show which method performs well due to privacy level and information loss. We found that each method has its limitations and advantages, so we need to combine the advantages of this method and propose a new approach that is effective in terms of privacy and data utility.

1.1 Our Main Contributions Are as Follows

1) We present a privacy framework for 1:M data publishing that makes use of (k, l)-diversity, as well as analytical results for the proposed model.By enforcing k-anonymity on the SA fingerprint and l-diversity on each equivalence class, (k, l)-diversity can protect QID and SA information during 1:M data publishing.
2) We develop an efficient method for 1: M-Generalization consisting of the combination of two algorithms: the Mondrian top-down greedy algorithm and the 1:M-Generalization algorithm, based on the (k, l)-diversity model, and compare the result between these algorithms according to information loss and efficiency according to execution time
3) applying the different methods to the two real-world datasets and comparing them to show which method performs well due to privacy level and information loss.

2 Related Work

Over the past several years, different techniques have been proposed to preserve privacy, and there are several techniques for privacy-preserving data publishing (PPDP) and privacy preserving data mining (PPDM).

We briefly discuss these methods and algorithms. Following are some novel works on privacy preservation. Since the introduction of the k-anonymity algorithm that was implemented by Sweeney in [1, 9, 10], many anonymization techniques [9, 13] and privacy models [11, 12] have been developed to prevent disclosure and privacy leakage during data publishing. To make sure that no record in the dataset can be distinguished from any other (k-1) record on QIDs, the author created k-anonymity. In [12], they discovered a k-anonymity barrier: a k-anonymized dataset is vulnerable to background

knowledge and homogeneity attacks. They developed the L-diversity model with SA diversity chains to enhance privacy protection. In [11], they investigated the similarity and skewness attacks on l-diversity and proposed a t-closeness model with a distribution constraint to protect privacy. However, t-closeness cannot effectively protect the privacy of infrequent values, which are more vulnerable to privacy disclosure. In [18], they proposed the Mondrian multidimensional k-anonymization algorithm, which is based on top-down greedy approximation. In [24], they proposed a technique called Anatomy, which anonymizes microdata sets by damaging a relationship between QID and SA characteristics. But because Anatomy divulges specific QID values, it is susceptible to presence attacks. In [29], a proposed MNSACM method using clustering and multi-sensitive bucketization for anonymizing the datasets that contain both numerical and categorical sensitive attributes is described, but it is limited to numerical sensitive attributes only and has not been implemented on real datasets. Not suitable for incremental data sets. in [28] To anonymize set-valued data, a local differential privacy technique called LDPMiner was created. These techniques are only useful for publishing results; they cannot be used to publish secure data sets for sharing.

Privacy model (technique)	evaluation	Disclosure of Privacy	Utility of Data	references
Slicing	designed to handle high-dimensional data, although slicing can return the original tuples after random tuple permutation if separate tuples have the same QIDs and SAs	Attacks of Skewness and Similarity	loss of information	[33]
Anatomization + Slicing	The method takes a long time to complete and generates numerous tables. The solution is extremely intricate	Knowledge Attack Regarding Demography	loss of information	[26]
SLOMS	The suggested approach removed the link between MSA. As a result of generalization, it issued numerous tables with information loss	Knowledge Attack Regarding Demography	loss of information	[34]

(*continued*)

(*continued*)

Privacy model (technique)	evaluation	Disclosure of Privacy	Utility of Data	references
MSA (α, l)	Utilized suppression and generalization with anatomy, which reduced utility	----------	High information loss	[35]
LKC- privacy + slicing	For dynamic MSA data posting, the KC slice technique was offered; however, the authors did not give examples of numerous releases, thus I have just covered the entire technique for a single release	----------	High information loss	[27]
(p, k)-Anonymity + Angel	The suggested method keeps MSA's healthcare microdata private. The (p, k) angelization heuristic algorithm serves as its foundation This method, however, is susceptible to sensitive attribute correlation attacks and MSA quasi attribute assaults	Attacks with MSA correlation	Low Information loss	[15]
(p,l)-Angelization	The suggested method, which combines the 1:M dataset and the MSA dataset, protects the privacy of healthcare microdata	generalization correlation attacks	Low Information loss	[36]

(*continued*)

(*continued*)

Privacy model (technique)	evaluation	Disclosure of Privacy	Utility of Data	references
G-Model	G-model protects against gender-specific SA attacks by maintaining separate groups and caches of male and female SAs. The G-model also stays away from generalization	Semantic correlation attacks	fail to provide optimal privacy protection	[37]

3 Problem Setting and Privacy Model

IOT devices generate a large amount of data, and this data may contain 1: m records or MSAs for each individual that lead to new privacy leaks due to different attacks such as similarity attacks, membership attacks, background knowledge attacks, and linking attacks. We consider 1: M-MSA dataset that can be faced many problems.

Two main problems are:

Problem 1: The majority of privacy models and techniques, such as K-anonymity, L-diversity, anatomization, and slicing, don't take the semantic relationship between sensitive values into account and are unable to process multiple attributes with various sensitivity requirements when simulating real-world privacy requirements for data publishing, which results in privacy disclosure.

Problem 2: (failure of Privacy model). Due to a dataset's many instances of a single individual, applying conventional 1:1 privacy methods to 1: M datasets and MSA may result in privacy exposure issues. K-anonymity, L-diversity, and SLOMS are examples of models. The key vulnerability information is that if an adversary is successful in identifying a single sensitive attribute, further sensitive qualities can be detected through co-relation.as the current framework falls short of ensuring data utility and protecting privacy for a number of sensitive characteristics.

We use the heart disease dataset to perform a comparative study between the different anonymization techniques for multiple sensitive attribute datasets. The dataset INFORMS consists of several attributes, such as month of birth, years of education, year of birth, marital status, race, income, diagnosis codes, sex, and disease. Heart Disease is a dataset with several attributes, including (gender: gender (1 = male; 0 = female), cp: chest pain type: Value 1: typical angina Value 2: atypical angina Value 3: non anginal pain Value 4: asymptomatic, trestbps: resting blood pressure (in mm Hg on admission to the hospital), fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false), thalach: maximum heart rate achieved; resting electrocardiographic results). The original microdata as in Table 1.

4 Anonymization Algorithms and Methods

In this section, we discuss two algorithms and three methods to show which one performs well in terms of information loss, data utility, and execution time by comparing them.

Mondrian Algorithm on IT

The Mondrian algorithm is a popular one for relational data anonymization [18]. It can successfully anonymize QID in a top-down manner. Mondrian is the fastest local recording algorithm while retaining good data utility; categorical attributes have no generalization hierarchies. This operation results in lower information loss, but worse semantic results.

The main phases of Mondrian algorithm:

1. Use kd-tree to divide the raw dataset into k groups. Every group in a K-group consists of at least k records.
2. Ensure that each k-group has the same QID by generalizing each k-group.

5 1: M-Generalization Algorithm:

1M_Generalization is an anonymization algorithm for 1: M dataset. It contains two sub-algorithms: Mondrian (for the relational part) and Partition (for the transactional part). Both of them are straight forward based on (k, l)-diversity model by enforcing k-anonymity on the SA fingerprint and l -diversity on each equivalence class, (k, l)-diversity can protect QID and SA information during 1: M data publishing. This algorithm performs data partitioning twice. Perform SA anonymization first, followed by QID generalization. Use the Mondrian [18] for QID generalization and the Partition [22]. For SA anonymization. Following are the main stages of 1:M-generalization:

Step 1: Transformation convert 1:M microdata first to 1:1 microdata.
Step 2: SA fingerprint anonymization.
Step 3: QID anonymization and SA diversity.

Algorithm 1: Partition
in relation to the partition with SA fingerprints.
Partition (partition, k)
Adding the partition to the global return list if the partition cannot be separated;
else //
The following steps are taken: / choose a node with the maximum information gain split_Node choose node (partition); // distribute records to sub partitions
Sub_ Partitions ← data distribution (partition, split_Node);
// deal with subp_artitions that have fewer than k records
balance partitions (sub Partitions);
for sub_ Partition in sub _Partitions do
Partition (sub _Partition, k);

Algorithm 2: Mondrian for data in 1: M Dataset
Mondrian (partition, l)
Add the partition to the global return list if the part's ion cannot be divided;
else
/* It is best to choose the property with the widest (normaliZed) range of values. */
dm ← Select the attribute (partition);
When dm is a number,
threshold ← Choose Threshold (dm, partition);
lHS← {t ∈ partition: T [dm] ≤threshold};
rHS← {t ∈ partition: T [dm] > threshold}; sub Partition ← {lHS} _ {rHS};
else
split_Node ← split (partition, dm);
sub_ Partitions ← share information (partition, split Node);
for sub Partition in sub Partitions do
Mondrian (sub _Partition, l) ;

Fig. 1. Anonymization Algorithms (Partition, Mondrian)

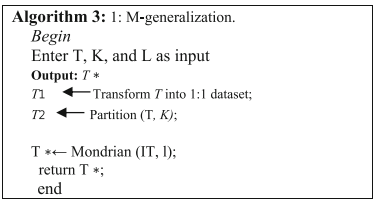

Fig. 2. 1: M-generalization Algorithm

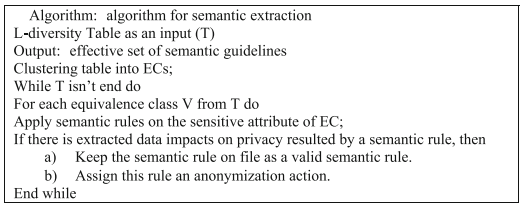

Fig. 3. Semantic Anonymization algorithms [24]

A. **Semantic Anonymization Method**

The proposed method (Fig. 1) is based on the L-diversity approach and semantic extraction techniques. The semantic extraction is based on semantic rules that can be assigned by the owner of the data that will be published [24]. Describe the relationship between sensitive values. The semantic rules that produce data, which have an impact on privacy and lead to similarity attacks, are called "effective semantic rules."

The anonymizer determines the important pieces of information that need to be anonymized and then selects the effective semantic rules that produce these pieces of information. The approach consists of two main parts: Data extraction using effective semantic principles to determine the rules of anonymization Table 2 displays the ECs, the applicable semantic rule for them, and the suggested action for anonymization in accordance with the values of sensitive characteristics.

Our suggested rules:

In INFORMS dataset:

Rule 1: If "the bronchitis, lung cancer, cough, flu and bronchitis" then is Respiratory infection disease.
Rule 2: If "income with values 0 to 2000" then is very low salary.

In Heat disease dataset:

Rule 1: If "cp is typical angina or atypical angina" then is heart disease.
Rule 2: If "trestbps from 150 to 200" then is high resting blood pressure. Rule 3: If "fbs $>$ 120" then blood sugar disease. Rule 4: If "thalach from 170 to 202" then is heart disease. Rule 5: If "restecg ST $>$ .5" then is heart disease.

Table 3 demonstrates the ECs, the relevant semantic rule, and generates action for anonymization in accordance with the semantic similarity between sensitive attribute values, the red cells refer to the effective semantic rules (described above) that produce data, which have an impact on the privacy and lead to privacy disclosure, to solve this issue, we applying the anonymization, the data is shown in Table 4. It is noticed that the EC which has QI {[30, 40], F} is merged with EC which has QI {[40, 50], M} into one EC where become has QI {[30, 50], *} and contain all sensitive values that were in the two EC. Therefore, prevent attacker from discloser the privacy of SA (Figs. 2 and 3).

Table 1. Original microdata (the heart disease dataset before anonymization)

Age	Sex	cp	trestbps	fbs	thalach	Restecg
28	M	typical angina	145	108	150	0
29	M	asymptomatic	130	101	162	1
21	M	atypical angina	120	110	148	0
41	M	non-anginal pain	140	105	153	0
50	M	atypical angina	138	110	151	1
48	M	asymptomatic	120	110	114	1
36	F	typical angina	160	125	187	1
37	F	atypical angina	150	130	172	1
30	F	atypical angina	172	130	178	1

Table 2. Entropy 3_Diversity (the heart disease dataset after 3_Diversity)

Age	Sex	cp	trestbps	fbs	thalach	Restecg
[20, 30]	M	typical angina	145	108	150	0
[20, 30]	M	asymptomatic	130	101	162	1
[20, 30]	M	atypical angina	120	110	148	0
[30, 40]	F	atypical angina	172	130	178	1
[30, 40]	F	typical angina	160	125	187	1
[30, 40]	F	atypical angina	150	130	172	1
[40, 50]	M	non-anginal pain	140	105	153	0
[40, 50]	M	asymptomatic	120	110	114	1
[40, 50]	M	atypical angina	138	110	151	1

Table 3. The effective semantic rules and anonymization action

Age	**Sex**	**cp**	**trestbps**	**fbs**	**thalach**	**Restecg**	**Rules**	**action**
[20,30]	M	typical angina	145	108	150	0	.....	.....
[20,30]	M	asymptomatic	130	101	162	1	.....	.....
[20,30]	M	atypical angina	120	110	148	0	.....	.....
[30,40]	F	atypical angina	172	130	178	1	heart disease blood sugar disease high resting blood pressure	Generalization, Adding +10 Suppress Sex attributes
[30,40]	F	typical angina	160	125	187	1		
[30,40]	F	atypical angina	150	130	172	1		
[40,50]	M	non-anginal pain	140	105	153	0	.....	.....
[40,50]	M	asymptomatic	120	110	114	1	.....	.....
[40,50]	M	atypical angina	138	110	151	1	.....	.....

Table 4. 3_Diversity after anonymization

Age	Sex	cp	trestbps	fbs	thalach	Restecg
[20, 30]	M	typical angina	145	108	150	0
[20, 30]	M	asymptomatic	130	101	162	1
[20, 30]	M	atypical angina	120	110	148	0
[30,50]	*	atypical angina	172	130	178	1
[30,50]	*	typical angina	160	125	187	1
[30,50]	*	atypical angina	150	130	172	1
[30,50]	*	non-anginal pain	140	105	153	0
[30,50]	*	asymptomatic	120	110	114	1
[30,50]	*	atypical angina	138	110	151	1

B. **Rating Privacy Preservation method**

Rating method (Fig. 4) used to disseminate sensitive data The AT (Attribute Table) and IDT (ID Table) tables for rating releases are based on various sensitivity coefficients for various qualities. This approach preserves many microdata correlations while simultaneously safeguarding the privacy of numerous sensitive variables. [25]. Increase the usefulness of released data while meeting the privacy requirements for a variety of sensitive features. This technique calculates the (sensitivity coefficient SCi).

Table 5. Miodata used for rating method

Age	Sex	cp	trestbps	fbs	thalach	Restecg
28	M	typical angina	145	108	150	0
29	M	asymptomatic	130	101	162	1
21	M	atypical angina	120	110	148	0
30	M	atypical angina	172	130	178	1
36	M	non-anginal pain	160	125	187	0
37	F	atypical angina	150	130	172	1
41	F	non-anginal pain	140	105	153	0
48	F	asymptomatic	120	110	114	1
56	F	atypical angina	138	110	151	1
65	F	asymptomatic	150	105	148	0

Algorithm: *AiIDj*-Creation
input: T with SC1, SC2, SCd+1
Results: AiIDj
1: j=0 *AiIDj=Ø*
2: For each property Ai, multiply by $(1 \le i \le d + 1)$
3: based on their values, hash the Aitk in Ai (each bin according to value)
/* The group-creation process is in lines 4–9. */
4: so long as at least SCi non-empty hash buckets exist;
/* lines 5 to 9 create a new AiIDj. */
5: j=j+1; *AiIDj = Ø*
6: S is the collection of SCi buckets with the largest size at the moment.
7: for each bucket in S
8: remove a random tuple Aitk from the collection
9: *AiIDj = AiIDj U {Aitk}*
/* The step of residue assignment is lines 10 to 13. */
10: for each non-empty bucket
/* There is just one value for this bucket; see Lemma 3. */
11: *Ait'k* = the bucket's sole remaining asset
12: *S'* = the collection of AiIDj without the Ait'k
13: pick a random AiIDj in S' to receive Ait'k.

Fig. 4. Rating method [25]

Table 6. Rating publishes AT with SC (3; 2; 2; 2; 2; 2; 2) for microdata in Table 5.

tk	Age (A1)	Sex(A2)	cp(A3)	Trestbps (A4)	Fbs (A5)	Thalach (A6)	Restecg (A7)
t1	A1ID1	A2ID1	A3ID2	A4ID1	A5ID1	A6ID1	A7ID1
t2	A1ID1	A2ID2	A3ID2	A4ID2	A5ID2	A6ID2	A7ID2
t3	A1ID1	A2ID3	A3ID1	A4ID3	A5ID3	A6ID3	A7ID3
t4	A1ID2	A2ID4	A3ID3	A4ID4	A5ID4	A6ID4	A7ID4
t5	A1ID2	A2ID5	A3ID5	A4ID5	A5ID5	A6ID5	A7ID5
t6	A1ID2	A2ID1	A3ID4	A4ID1	A5ID1	A6ID1	A7ID1
t7	A1ID3	A2ID2	A3ID4	A4ID2	A5ID2	A6ID2	A7ID2
t8	A1ID3	A2ID3	A3ID3	A4ID3	A5ID4	A6ID3	A7ID3
t9	A1ID3	A2ID4	A3ID5	A4ID5	A5ID5	A6ID4	A7ID5
t10	A1ID3	A2ID5	A3ID1	A4ID4	A5ID3	A6ID5	A7ID4

C. (p, k) angelization for MSAS data publication with weight calculation

Since each bucket contains entries that fall within the p categories, each bucket partitioning complies with the (p, k)-anonymity principle. This technique (Angelization)

Table 7. For microdata in Table 5, rating publishes IDT with SC (3, 2; 2; 2; 2). (aty = atypical angina, Ty = typical angina, asy = asymptomatic, non = non-anginal pain)

IDj	Age (A1)	Sex(A2)	cp(A3)	Trestbps (A4)	Fbs (A5)	Thalach (A6)	Restecg (A7)
ID1	21,28,29	M,F	aty, asy	108,130	108,130	150,172	0,1
ID2	30,36,37	M,F	Ty, asy	130,140	101,105	162,153	0,1
ID3	41,48,56,65	M,F	aty, asy	120,120	110,105	148,114	0,1
ID4	-	M,F	aty,non	172,150	130,110	178,151	0,1
ID5	-	M,F	non, aty	160,138	125,110	187,148	0,1

comprises the bucket and batch partitioning pairs. At least k tuples are present in each bucket, where k is the minimum group size to prevent linking attacks (Tables 6 and 7).

This approach uses weighted measurements of SA (sensitive attributes), as all of these sensitive attributes have different levels of sensitivity or weight. The weights of all sensitive attributes are calculated for the purpose of identifying the level of sensitivity of each attribute. Let W = {w1, w2,……. wd} [15] w1 is the weight of s1, w2 is the weight of s2, as shown in Table 8 to indicate the set of weights assigned to each sensitive attribute., and so on. There are two published tables as a result (a generalized table and a sensitive batch table), as shown in Tables 9 and 10.

Table 8. Weight calculation

Sensitive attributes	Identified by	Dependency	Weightage
S1 = cp: chest pain type	S2,S3,S4	3	3
S2 = trestbps: resting blood pressure	---------	0	0
S3 = restecg: resting electrocardiographic results	S2,S5	2	2
S4 = thalach: maximum heart rate achieved	S1,S2,S5	3	3
S5 = fbs: fasting blood sugar	S2	1	1

A single table contains numerous, highly correlated attributes. For example, {cp, thalach} and {restecg, trestbps, fbs} as shown in Table 11 and Table 12. In order to ensure l-diversity in each bucket, horizontal partitioning is used (Fig. 5).

Using (p, k) angelization algorithm which the output consists of Generalized table and Sensitive batch Table (Table 13 and Table 14).

Table 9. QUASI-TABLE (QIT) (publish all quasi identifiers for each individual)

Tuple id	Age	Sex
P1	**28**	**M**
P2	29	M
P3	21	M
P4	30	F
P5	36	F
P6	37	F
P7	41	M
P8	48	M
P9	46	M

Table 10. Sensitive attribute table (ST) (publish all sensitive attributes for each individual)

cp	trestbps	fbs	thalach	Restecg
typical angina	145	108	150	0
asymptomatic	130	101	162	1
atypical angina	120	110	148	0
atypical angina	172	130	178	1
typical angina	160	125	187	0
atypical angina	150	130	172	1
non-anginal pain	140	105	153	0
asymptomatic	145	115	114	1
atypical angina	138	110	151	1

Table 11. Sliced sensitive attributes (CP, THALACH)

Tuple ID	cp	thalach	Group
P1	typical angina	150	1
P2	asymptomatic	162	
P3	atypical angina	148	
P4	atypical angina	178	2
P7	non-anginal pain	187	
P9	atypical angina	151	
P5	typical angina	153	3
P8	asymptomatic	114	
P6	atypical angina	172	

Table 12. Sliced sensitive attributes (RESTEC, TRESTBPS, FBS)

Tuple ID	restecg	trestbps	fbs	Group
P1	0	145	108	1
P2	1	130	101	
P5	0	160	125	
P4	1	172	130	2
P3	0	120	110	
P6	1	150	130	
P7	0	140	105	3
P8	1	145	110	
P9	1	138	115	

Input:
1: micro dataset
A) Explicit identifier(E)
B) quasi – identifier attributes
C) sensitive attributes
2: EX: **(External factor)**
Output:
A) Generalized Table
B) Sensitive Batch Table

Fig. 5. "(p, K) – Angelization": [15]

Table 13. Generalized table

Age	Sex	Batch ID
[20, 30]	Person	**1**
[20, 30	Person	
[20, 30	Person	
[30,38]	Person	**2**
[30,38]	Person	
[38,45]	Person	**3**
[38,45]	Person	
[45,50]	Person	**4**
[45,50]	Person	

Table 14. Sensitive batch table (SBT)

cp	trestbps	fbs	thalach	Restecg	Batch ID
typical angina, asymptomatic	145, 130	108, 101	150, 162	0,1	1
non-anginal pain, atypical angina	140, 172	105, 130	153, 178	0,1	2
atypical angina, asymptomatic	120, 145	110, 115	148,115	0,1	3
typical angina, atypical angina	160, 138	125, 110	187, 151	0,1	4

6 Experiments and Analysis Result

We use different algorithms and methods to anonymize the dataset and show the effects of each method due to data privacy, information loss, and computational efficiency. Execution time and the quality of anonymized data are measured. to compare 1: M-generalization algorithm, Mondrian algorithm. We use two actual datasets. (Heart disease, INFORMS) from UCI repository. By contrasting these techniques with l-diversity, we assess the effectiveness of these techniques., and we apply the method of semantic anonymization to solve the problem of similarity attack. In Informs, we select (sex, maternal status, and education) as QIA and (income, disease) as SA. We select $l = 5$ and $k = 10$ as the default values. The Heat Disease dataset: this dataset contains 75 attributes. We have taken 2 quasi-identifier attributes (sex and age) and (cp: chest pain type; trestbps: resting blood pressure, restecg: resting electrocardiographic results; thalach: maximum heart rate achieved; fbs: fasting blood sugar) as SAs.

6.1 Information Loss and Data Utility

We employed NCP (Normalized Certainty Penalty) [19] to quantify the information loss caused by anonymization. We calculate QIDNCP and SANCP with different parameters and present the results. We observed that QID–NCP increases when l grows because a larger l necessitates a larger size for each EC, which implies that QID values can be more broadly generalized; however, it provides more privacy guarantee because the criterion for l-diversity offers superior protection to the requirement for k-anonymity when l is larger than k. Meanwhile, both SA–NCP and QID–NCP of 1: M-generalization will increase slightly when k increases, and in other algorithms the NCP will increase slightly when k increases see Table 15.

The outcome demonstrates that the rating method enables data analysis that is substantially more efficient than l-diversity [23]. For categorization, rating performs better than Actually Rating performance for classification is comparable to that of microdata. In other words, the information about the original data is not significantly lost during the rating process. The (p, k)-Angelization strategy guards against non-membership attacks and demographic attacks on the rating method. Also decreases information loss, increasing the usefulness of the information that has been released publicly. But have more execution time due to the complicated process of weight calculation and the (p, k)-Angelization publication of numerous tables. The more release tables there are, the longer it takes to execute.

6.2 Efficiency

We calculate the entire execution time of our methods to evaluate their effectiveness (except for the pre-processing). Specifically, we tested our algorithms on several datasets (l and k). For $k = 10$, $1 = 5$ as a default parameter, 1:M-generalization typically takes less than 28 s to execute. When k increases, the overall execution time of the 1:M-generalization is marginally reduced. This is due to the fact that a higher k suggests fewer partition splits. On the other hand, running time does not appear to change much with parameter l. The average execution time of the Mondrian algorithm is less than 20 s, and Mondrian_l_diversity is less than 10 s; see Table 16.

We tested the effect of the similarity attack on data after semantic anonymization. The results of the proposed method show enhancement in terms of privacy but decrease the utility of data due to information loss as a result of generalization and suppression in the anonymization process. We found that the balancing point between utility and privacy depends on the dataset and value of L. As the number of quasi-identifiers increases, the balancing point shifts downward and balances between Utility and privacy occur at a higher value of L, as shown in Fig. 6**.** Which plots the performance curves of anonymous data information loss over various L. The execution time of the (p, k)-Angelization is compared with the rating technique. The execution time has been measured while increasing the number of records. Since the rating technique publishes multiple tables, its execution time is larger than the (p, k)-Angelization as represented in Fig. 7. It has been calculated by varying the number of sensitive attributes from 1 to 6. The results are shown in Figs. 7 and 8 (Table 17).

Table 15. Evaluation of information loss in percentage

Algorithm name	k-value	l-value	NCP-percentage
1:M-generalization	10 20	5 5	QID–NCP = 11.68% SA–NCP = 6.70% QID NCP = 11.66% SA–NCP = 7.87%
Mondrian algorithm	10 20	……	NCP = 12% NCP = 15.03%

Table 16. Evaluation of efficiency based on execution time

Algorithm name	K -value	L-value	Actual Execution time
1:M-generalization	10 20	5 5	21.9 s 18.84 s
Mondrian algorithm	10 20	…… …	2 s 1. 2 s

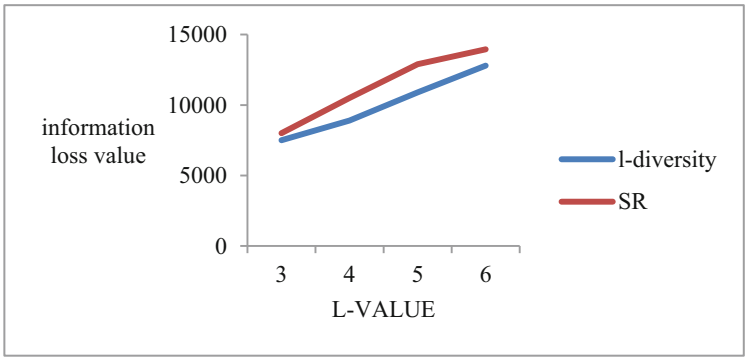

Fig. 6. Information loss for various l.

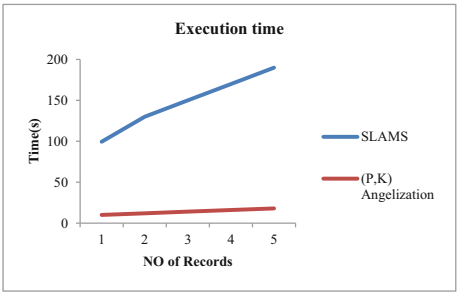

Fig. 7. Execution time when changing No of records

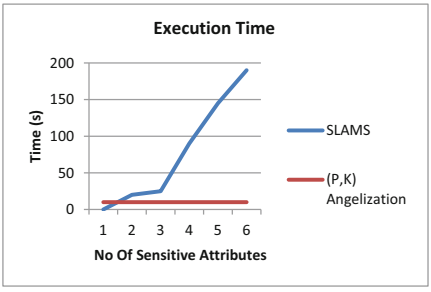

Fig. 8. Execution time when changing No of SA

As a result, the data publisher must strike a balance between the level of privacy required and the utility of the data (to minimize information loss). So we need to migrate

Table 17. Comparison between the different methods

Method or technique	Privacy requirement	Utility	Time execution	Method used for measurement of utility
Semantic rules	More enhancements to the term privacy. Than other methods Reduces of similarity attack	More information loss than other methods due to generalization and suppression decrease the utility	Less time Execution	(Entropy, recursive (c, L))-diversity with different values for L using the generalized loss metric, Discernibility Metric (query answering)
Rating	Achieved the required level of privacy but less than other two methods. Compared to l-diversity, that ranking has nearly a hundred times less ARE solves the "curse of dimensionality" issue	Low information loss High utility More effective data analysis. The performance the performance of rating is close to the performance of microdata for classification	Less time Execution	Classification measurement and average relative error measurement (ARE) (mining process)
(p,k) angelization	Provides enhanced privacy that Achieved the most required level of privacy Reduces of correlation attacks and prevents demographic attack and non-membership attacks	Low information loss than other two methods	more time Execution due to the process of weight calculation	Demographic error (DE) KL-divergence and DCP (Discernibility penalty) (query answering)

or combine the advantages of these methods and propose a new method that can achieve the target of balanced privacy and utility. From this comparison, we found that the 1: m-generalization algorithm and the (p, k) angelization method perform well in terms of information loss and data utility.

7 Conclusion and Future Work

This paper stands out a study and concept of multiple sensitive attributes (MSAs). 1: M records of an individual and MSAs in microdata separately is a topic that the majority of earlier studies address; the generalization of QID or MSAs, which results in significant information loss and low data value, are two major drawbacks of this strategy. While MSA privacy models make an effort to prevent privacy exposures, such strategies are either overly complex or lacking in some crucial components. Consequently, there is a need for an efficient solution for 1: M with MSA that strikes a balance between data utility and privacy. We apply different privacy methods to address this problem. Overall, experiments with real-world datasets point out that these methods outperform the most recent developments in terms of privacy, information loss, and execution time. Our work applied on MSAs in the future we seek to handle the dataset that contain also multiple records for each individual with MSA. Additionally, this approach assumes a static dataset will only be published once, as opposed to a dynamic dataset being repeatedly republished. It is yet unclear how 1: M with MSAs works across different re-publications of a dynamic dataset. Future research will also examine similar repeated publications of dynamic data rather than static data. (data streams).

References

1. Latanya, S.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(05), 557–570 (2002)
2. Yaseen, S., Saba, T., Anjum, A.: Improved generalization for secure data publishing. IEEE Access **6**, 27156–27165 (2018)
3. Komishani, E.G., Abadi, M., Deldar, F.: PPTD: preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. Knowl.-Based Syst. **94**, 43–59 (2016). https://doi.org/10.1016/j.knosys.2015.11.007
4. Wang, J., Du, K., Luo, X., Li, X.: Two privacy-preserving approaches for data publishing with identity reservation. Knowl. Inf. Syst. **60**(2), 1039–1080 (2018). https://doi.org/10.1007/s10115-018-1237-3
5. Wang, R., Zhu, Y., Chen, T.S., Chang, C.C.: Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness. J. Comput. Sci. Technol. **33**(6), 1231–1242 (2018)
6. Tao, Y., Tong, Y., Tan, S., Tang, S., Yang, D.: Protecting the publishing identity in multiple tuples. In: Atluri, V. (ed.) DBSec 2008. LNCS, vol. 5094, pp. 205–218. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-70567-3_16
7. Poulis, G., Loukides, G., Skiadopoulos, S., Divanis, A.J.: Anonymizing data with relational and transaction attributes. In: European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg (2013)
8. Anjum, A., Ahmed, T., Khan, A., Ahmad, N.: Privacy preserving data by conceptualizing smart cities using MIDR-Angelization. Sustain. Cities Soc. **40**, 326–334 (2018)
9. Ting, Y., Jajodia, S.: Secure data management in decentralized systems, vol. 33. Springer Science & Business Media (2007)
10. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity In:. IEEE 23rd International Conference on Data Engineering (2007)
11. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: 22nd International Conference on Data Engineering (ICDE'06). IEEE (2006)

12. Ercan Nergiz, M., Clifton, C.: Thoughts on k-anonymization. Data Knowl. Eng. **63**(3), 622–645 (2007). https://doi.org/10.1016/j.datak.2007.03.009
13. Jabeen, F., Hamid, Z., Wadood, A., Ghouzali, S.: Enhanced architecture for privacy preserving data integration in a medical research environment. IEEE Access **5**, 13308–13326 (2017)
14. Anjum, A., Ahmad, N., Malik, S.U.R., Zubair, S., Shahzad, B.: An efficient approach for publishing microdata for multiple sensitive attributes. The Journal of Supercomputing **74**(10), 5127–5155 (2018). https://doi.org/10.1007/s11227-018-2390-x
15. Lee, H., Kim, S., Kim, J.W., Chung, Y.D.: Utility-preserving anonymization for health data publishing. BMC Med. Inform. Decis. Mak. **17**, 104 (2017)
16. Majeed, A.: Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data. J. King Saud Univ.-Comput. Inform. Sci. **31**, 426-435 (2018)
17. LeFevre, K., De Witt, D.J., Ramakrishnan, R.: Mondrian Multidimensional k-anonymity, vol. 6. ICDE (2006)
18. Xu, J.,Wang, W., Pei, J., Wang, X.: Utility-based anonymization using local recoding. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (2006)
19. Abdelhameed, S.A., Moussa, S.M., Khalifa, M.E.: Privacy-preserving tabular data publishing: a comprehensive evaluation from web to cloud. Comput. Secur. **72**, 74–95 (2018)
20. Anjum, A., Ahmad, N., Raza, B.: An efficient privacy mechanism for electronic health records. Comput. Secur. **72**, 196–211 (2018)
21. He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local generalization. Proc. VLDB Endow. **2**(1), 934–945 (2009)
22. Lu, H., Setiono, R., Liu, H.: Effective data mining using neural networks. IEEE Trans. Knowl. Data Eng. **8**(6), 957–961 (1996)
23. Mubark, A.A., Elabd, E., Abdulkader, H.: Semantic anonymization in publishing categorical sensitive attributes. In: 8th International Conference on Knowledge and Smart Technology (KST). IEEE (2016)
24. Liu, J., Luo, J., Huang, J.Z.: Rating: privacy preservation for multiple attributes with different sensitivity requirements. In: IEEE 11th International Conference on Data Mining Workshops. IEEE (2011)
25. Susan, V.S., Christopher, T.: Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes. Springerplus **5**(1), 1–21 (2016)
26. Onashoga, S.A., Bamiro, B.A., Akinwale, A.T., Oguntuase. J.A.: KC-Slice: A dynamic privacy-preserving data publishing technique for multisensitive attributes. Inform. Secur. J.: A Global Perspect. **26**(3), 121–135 (2017)
27. Zaman, A.N.K., Obimbo, C., Dara, R.A.: An improved data sanitization algorithm for privacy preserving medical data publishing. In: Mouhoub, M., Langlais, P. (eds.) AI 2017. LNCS (LNAI), vol. 10233, pp. 64–70. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57351-9_8
28. Liu, Q., Shen, H., Sang, Y.: A privacy-preserving data publishing method for multiple numerical sensitive attributes via clustering and multi-sensitive Bucketization. In: Sixth International Symposium on Parallel Architectures, Algorithms and Programming. IEEE (2014)
29. He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local gen-eralization. In: Proceedings of the 35th International Conference on Very Large Data Bases(VLDB), VLDB Endowment (2009)
30. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering(ICDE), p. 25. IEEE Computer Society (2006)
31. Li, T., Li, N., Zhang, J., Molloy, I.: Slicing: a new approach for privacy preserving data publishing. IEEE Trans. Knowl. Data Eng. **24**(3), 561–574 (2012)

32. Han, J., Luo, F., Lu, J., Peng, H.: SLOMS: a privacy preserving data publishing method for multiple sensitive attributes microdata. JSW **8**(12), 3096–3104 (2013)
33. Abdalaal, A., Nergiz, M.E., Saygin, Y.: Privacy-preserving publishing of opinion polls. Comput. Secur. **37**, 143–154 (2013)
34. Kanwal, T., Anjum, A.: Privacy-preserving model and generalization correlation attacks for 1: M data with multiple sensitive attributes. Inf. Sci. **488**, 238–256 (2019)
35. Albulayhi, Kh., Tosic, P.T., Sheldon, F.T.: G-model: a novel approach to privacy-preserving 1:M microdata publication. In: 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom). New York, NY, USA (2020)