f-slip: An efficient Privacy-preserving data publishing framework for 1:M microdata with multiple sensitive attributes.

Jayapradha. J[1], Prakash. M[2]

Department of Computer Science and Engineering, College of Engineering and Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, 603203, Kanchipuram, Chennai, TN, India, jayapraj@srmist.edu.in

Department of Computer Science and Engineering, College of Engineering and Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, 603203, Kanchipuram, Chennai, TN, India, prakashm2@srmist.edu.in

Abstract

Privacy of the individuals plays a vital role when a dataset is disclosed in public. Privacy-preserving data publishing is a process of releasing the anonymized dataset for various purposes of analysis and research. The data to be published contain several sensitive attributes such as diseases, salary, symptoms, etc. Earlier, researchers have dealt with datasets considering it would contain only one record for an individual [1:1 dataset], which is uncompromising in various applications. Later, many researchers concentrate on the dataset, where an individual has multiple records [1:M dataset]. In the paper, a model f-slip was proposed that can address the various attacks such as Background Knowledge (bk) attack, Multiple Sensitive attribute correlation attack (MSAcorr), Quasi-identifier correlation attack(QIcorr), Non-membership correlation attack(NMcorr) and Membership correlation attack(Mcorr) in 1:M dataset and the solutions for the attacks. In *f*-slip, the anatomization was performed to divide the table into two subtables consisting of i) quasi-identifier and ii) sensitive attributes. The correlation of sensitive attributes is computed to anonymize the sensitive attributes without breaking the linking relationship. Further, the quasi-identifier table was divided and k-anonymity was implemented on it. An efficient anonymization technique, frequency-slicing (f-slicing), was also developed to anonymize the sensitive attributes. The *f*-slip model is consistent as the number of records increases. Extensive experiments were performed on a real-world dataset Informs and proved that the *f*-slip model outstrips the state-of-the-art techniques in terms of utility loss, efficiency and also acquires an optimal balance between privacy and utility.

Keywords: 1:M dataset, Privacy-Preserving, Anatomization, k-anonymity, *f*-slicing, *f*-slip.

1. Introduction

Various organizations and institutions publish their data for research, analysis purposes, policy and decision making to make the data available for public and private sectors. Due to the increase in digital transformation, Electronic Health Record (EHR) also increased enormously. EHR includes complete information of the patient's disease, symptoms, demographics, diagnosis codes, test reports, allergies, physicians and bill reports (Chu et al. 2021; Orna et al. 2020). The usage of Electronic Health Records was 18% in 2001 and increased to 72% in 2012 and expected to rise to 90% at the end of the era (Stephen and Michael 2016). The data released by the health sectors for the analysis and research purposes may hold the personal information of an individual such as explicit identifiers (e.g., name, SSN), quasi-identifiers (e.g., name, age, sex, race) and sensitive attributes (e.g., disease, symptoms, salary). Publishing such data with private and personal information leads to a privacy breach and thus, the individual's privacy is compromised. Later, the explicit identities were removed from the data before publishing, stating that the microdata is secured. It is well-known that just removing the explicit identifiers is insufficient and may lead to linking attacks. The adversaries can connect the quasi-identifier attributes with the available external sources to re-identify a particular individual. Around 87% of the population in the U.S. was identified from the published medical records using the quasi-identifier such as gender, zip code and date of birth (Latanya 2002). Several privacy algorithms and models were proposed to overcome the privacy breaches (Abdul and Sungchang 2020; Zhe et al. 2018; Rashad and Azhar 2018). Health sectors and organizations anonymize their microdata with the existing privacy algorithms and models to protect individuals' from various privacy breaches. The threats in health sectors were dealt with by any techniques that have been proposed by computer science communities and health informatics (Kristen et al. 2005). The anonymization techniques aim at privacy-preserved data publishing with strong privacy and less information loss (Ismail and Ammar 2020; Jinwen et al. 2020; Tehsin et al. 2021).

Researchers carried out their work in privacy-preserving data publishing earlier, assuming that the dataset has a single sensitive attribute. However, the dataset may have multiple sensitive attributes (MSA). In a real-time dataset, each individual has multiple records with multiple sensitive attributes. For example, a patient might have

visited a particular hospital numerous times for various diseases (e.g., hypertension, bronchitis, diabetes). Each time the patient visits the hospital for different diseases, the records will be inserted into the database (e.g., three visits for hypertension and a single visit for bronchitis and a couple of visits for diabetes). Hence, a patient can have multiple records with multiple sensitive attributes (MSA). Most of the researchers have not concentrated on the above scenarios. Thus, the work focuses on 1: M microdata with MSA for Electronic Health Records.

The paper has been organized as follows. The related works from various researches are discussed in section 2. The motivation and challenges are elaborated with the sample dataset in section 3. Section 4 elucidates the 1:M multiple sensitive attribute attacks with corresponding scenarios. The contribution and preliminaries with various definitions have conversed clearly in sections 5 and 6. The model *f*-slip is explained elaborately in step by step manner in section 7. Section 8 outlines the algorithms of various processes in *the f*-slip model with a clear description. Experimental evaluation and result analysis are explained and depicted through various graphs in section 9 and section 10 concludes the paper with future directions and its limitations.

2. Related Works

This section evaluates the privacy anonymization methods and models for publishing the multiple sensitive attributes and 1: M microdata. Various anonymization methods and models have been proposed for MSA. K-anonymity (Latanya 2002; Khaled and Fida 2008) was proposed to prevent the re-identification of the individuals. However, k-anonymity could not restrict sensitive attribute disclosure and does not defend against reverse attack. (α,k)-anonymity was proposed with k-anonymity as the base. The distribution frequency of each sensitive attribute in every equivalence class should not be greater than 1/α (Xiangwen et al. 2017). The distribution of sensitive attributes reduces the competence of anonymity, so the model l-diversity (Ashwin et al. 2006) was proposed considering the diversity of the sensitive attributes. Further considering the above problems, an algorithm t-closeness was proposed with the measure of "distance between the distribution of the sensitive attributes in the equivalence class and the sensitive attribute distribution in the whole table should not be greater than the threshold t" (Ninghui et al. 2007). t-closeness fails in protecting the privacy of the infrequent values. Thus, β-likeness was proposed with strong constraints to achieve good privacy (Jianneng and Panagiotis 2012). When the existing privacy models were applied to the incremental data releasing model, they lead to excess privacy leakage; thus, model m-signature was proposed (Junqing et al. 2018). However, the m-signature model is limited by its time complexity application. An anatomy technique, Sensitive Label Privacy Preservation with Anatomization technique (SLPPA) and (α,β,γ,δ) model were proposed to prevent the record linkage, table linkage, attribute linkage and probabilistic attacks. A metric "mean-square contingency coefficient" was applied to divide the table and avoid re-identification during anatomization. The (α,β,γ,δ) model was applied on two datasets named adult and census and was limited to a single sensitive attribute. The SLPPA comprises of two processes i) table division and ii) group division (Lin et al. 2021). The above models and algorithms were concentrating on 1:1 microdata and single sensitive attribute.

The main focus of anonymization is to transform the data to balance both privacy and loss of information. Various anonymization techniques have been designed for privacy-preserving data publishing with MSA. SLAMSA is a privacy preservation approach for MSA. The SLAMSA, using an anatomization technique that prevents the generalization of quasi-identifier attributes, leads to less information loss. As SLAMSA anatomizes the original table into multiple tables, it causes complexity and utility loss during the publishing of tables. SLAMSA is implemented on Cleveland Foundation Heart Disease and Hungarian Institute of Cardiology datasets; however, it is vulnerable to demographic attacks (Shyamala and Christopher 2016). The concept KC_i-Slice considers different thresholds for different sensitive attributes. It prevents similarity attacks by applying semantic l-diversity. The KC_i-Slice reduces the utility loss and enhances the privacy for multiple sensitive attributes and is tested on the Adult dataset (Lakshmipathi et al. 2018).

The distributional model was proposed by setting a threshold p–sensitive on multiple sensitive attributes. A set of rules have been fixed for the sensitive attribute values distribution. The sensitive attributes are categorized as primary sensitive attributes and contributory sensitive attributes. In the distributional model, the sensitive attributes are divided into sub-tables without following the anatomy concept. Also, the distribution model was not fixed; it can be changed according to the different models (Widodo and Wahyu 2018). The novel method called overlapped slicing with bucketization technique was proposed for privacy-preserving data publishing with multiple sensitive attributes. In overlapped slicing, sensitive attributes are anonymized by applying permutation in each bucket. The Discernibility value metric was used to measure the utility and a comparison was made with two different existing methods and tested on an adult dataset. The overlapped slicing model lagged in dissociating the relationship between quasi-identifier and sensitive attributes (Widodo et al. 2019). The privacy and security

level of each sensitive attribute differs according to the different requirements of sensitivity. L_{sl}-diversity model was proposed with three greedy algorithms named maximal-bucket first (MBF), maximal single-dimension-capacity first (MSDCF) and maximal multi-dimension-capacity first (MMDCF) algorithm. The above three algorithms helped greatly in reducing the information loss. However, there was a slight increase in time when there is an increase in the volume of data (Yuelei and Haiqi 2020).

An effective approach (p,k)-Angelization was proposed to anonymize the multiple sensitive attributes. The (p,k)-Angelization eradicates the background join, non-membership attacks and yields the balance between privacy and utility. The approach (p,k)-Angelization, made one-to-one correspondence between the quasi-identifier and sensitive attributes in the buckets(Adeel et al. 2018). The "(*c, k*)-anonymization" is an advancement of (p,k)-Angelization, which enhanced the one-to-one correspondence to one-to-many correspondence to provide improved privacy and increased utility. (*c, k*)-anonymization also thwarts the "fingerprint correlation attacks"(Razaullah et al. 2020). Both (p,k)-Angelization and (*c, k*)-anonymization can be applied only for 1:1 microdata. Various models have been proposed and implemented on multiple sensitive attributes (Wang et al. 2018; Jayapradha et al. 2020; Lin et al. 2017). The papers discussed till now have implemented their works on 1:1 microdata, considering that datasets have only one record per person.

Later, few researchers have paid concentration towards 1: M datasets. A method called (k,k^m)anonymous was proposed, and it leads to unexpected information distortion. (k,k^m)anonymous framed the 1:M problem as "multi-objective optimization problem" and handled both relational and transactional data (Poulis et al. 2013). A method (*k,l*)-diversity was proposed to address the disclosure risk on privacy-preserving data publishing 1:M dataset. An algorithm 1:M generalization was proposed. However, unfortunately, it fails to prevent information loss (Qiyuan et al. 2017). A hybrid method *l*-anatomy was proposed to ensure the privacy of individuals on 1:M datasets. Though *l*-anatomy performed well in terms of utility, the computational complexity increased and was limited to a single sensitive attribute (Adeel et al. 2018). A bidirectional personalized generalization model was proposed to satisfy the higher privacy and less utility loss for multi-record datasets. This model resists bi-directional chain attack by using a hierarchical generalization strategy. Though the model performed well, it was limited to a single sensitive attribute and leads to information loss due to generalization (Xinning et al. 2020). QIAB-IMSB algorithm was proposed to anonymize the set-valued dataset. Vertical partitioning has been performed to partition the table. In QIAB-IMSB algorithm, k-anonymity has been applied for quasi-identifier bucket and (k,l)-diversity for multiple sensitive attribute bucket. The algorithm resists a sensitive linking attack by adopting hierarchical generalization and the accuracy of the data was compared using classification models (Jayapradha and Prakash 2021). As per the survey, the widely used real-world datasets for the 1:M privacy-preserving data publishing are Informs and YouTube.

3. Motivation and Challenges

In a real-world scenario, the 1:M datasets are more than 1:1 datasets. Apart from health care, there are various domains where users can possess multiple records. The individual might post multiple pictures and statuses on the same account in a social network such as Facebook, Twitter, Foursquare, etc. Likewise, a person can purchase various items on different days with the same membership card in the supermarket. Only a few researchers looked into the above scenario in their work earlier. Later, many researchers took the problem in hand and developed various privacy models and anonymization algorithms. However, the models and algorithms were not able to resist several disclosures and attacks.

Consider the 1:M dataset, shown in Table 1. It is a sample dataset that comprises patients' records with multiple sensitive attributes. In Table 1, patients consist of multiple records with different disease codes. The patient Alan has two records in the dataset, with two different disease codes. The age, gender and zip code are quasi-identifier and salary, poverty, education and disease code are sensitive attributes. The patient data is in the form of relational table R_T = {EI, QI, SA}. In Table R_T each individual can have multiple 'tp' tuples (i.e.) $\{rt_1, rt_2 \cup rt_3, rt_4 \cup rt_5, rt_6 \cup rt_7, rt_8, rt_9 \cup rt_{10} \cup rt_{11}\}$ (see Table 1). Table R_T consists of explicit identifiers EI= $\{ei_1,ei_2,ei_3 \ldots ei_n\}$, quasi-identifiers QI=$\{qi_1,qi_2,qi_3 \ldots qi_s\}$ and sensitive attributes SA= $\{sa_1,sa_2,sa_3 \ldots sa_h\}$.The explicit identifiers are given just to identify the patient, whereas it will be removed during data publication. The quasi-identifier comprises the general information that can be associated with the publically available dataset to re-identify an individual. Sensitive attributes contain confidential information such that the individual does not want to disclose it to the public. Therefore, sensitive attributes need to be protected from intruders.

Table 1 1: M sample microdata.

	Explicit Identifier		Quasi-identifier			Sensitive Attributes			
	Unique_ ID	**Name**	**Age**	**Gender**	**Zip code**	**Salary**	**Poverty**	**Education**	**Disease code**
rt_1	1	Lisa	19	F	60,000	20,000	medium	Bachelors	401
rt_2	2	Alan	13	M	55,000	10,201	high	9th	V22
rt_3	2	Alan	13	M	55,000	10,201	high	9th	V90
rt_4	3	Dalia	17	F	70,000	17,258	medium	12th	724
rt_5	3	Dalia	17	F	70,000	17,258	medium	12th	408
rt_6	4	Helen	30	F	68,000	25,364	low	Master	276
rt_7	4	Helen	30	F	68,000	25,364	low	Master	402
rt_8	5	Tony	16	M	75,000	19,223	medium	11th	492
rt_9	6	Tom	32	M	56,000	21,012	low	Master	272
rt_{10}	6	Tom	32	M	56,000	21,012	low	Master	404
rt_{11}	6	Tom	32	M	56,000	21,012	low	Master	490

Table 2 2-anonymity on 1:M microdata

	Explicit Identifier		Quasi-identifier			Sensitive Attributes			
	Unique_ ID	**Name**	**Age**	**Gender**	**Zip code**	**Salary**	**Poverty**	**Education**	**Disease code**
rt_1	1	Lisa	[15-20]	F	[50,000-60000]	20,000	medium	Bachelors	401
rt_2	2	Alan	[15-20]	M	[50,000-60000]	10,201	high	9th	V22
rt_3	2	Alan	1*	M	[55,000-75000]	10,201	high	9th	V90
rt_4	3	Dalia	1*	F	[55,000-75000]	17,258	medium	12th	724
rt_5	3	Dalia	17	[M,F]	[60,000-70,000]	17,258	medium	12th	408
rt_6	4	Helen	30	[M,F]	[60,000-70,000]	25,364	low	Master	276
rt_7	4	Helen	30	F	68,000	25,364	low	Master	402
rt_8	5	Tony	16	M	75,000	19,223	medium	11th	492
rt_9	6	Tom	32	M	56,000	21,012	low	Master	272
rt_{10}	6	Tom	3*	*	5****	21,012	low	Master	404
rt_{11}	6	Tom	3*	*	5****	21,012	low	Master	490

3.1 Challenge 1(Failure of 1:1 privacy model on 1:M)

When the existing privacy models of the 1:1 dataset are applied to the 1:M dataset, it might cause privacy breaches due to multiple records for an individual. Privacy models designed for 1:1 datasets can no longer be applied on 1:M datasets. In Table 2, 2-anonymity has been applied to protect the data against various privacy breaches. However, the datasets are not well-protected. Though Liu, Dalia, Alan, Helen, Tony and Tom records are generalized by forming equivalence classes, the patients can be easily re-identified. If an intruder knows the quasi-identifier values of Alan (i.e.) 13, M, 55,000, then the intruder can quickly identify the sensitive values of Alan. Since only the first two equivalence classes of Table 2 suits the above criteria, thus the intruder can infer the values of the sensitive attributes of Alan with 100% confidence as his salary is 10,201, the poverty line is high, education is 9th and also the disease codes are <V22,V90>. In Table 2, the records of the individuals are not aggregated. As the individual records have the same quasi-identifier values with different sensitive attribute values, the intruder can easily re-identify a specific person and infer an individual's complete information. In Table 2, the unique id (U_ID) of Alan and Dalia is in group 2. After exploring Table 2, the intruder can infer that U_ID 2 is generalized in groups 1 & 2; thus, the patient U_ID 3 quasi-identifier values can be inferred from U_ID 2 QI's values (i.e.) U_ID 3 should be < 20, female and zip code in the range of 60,000-70,000. In Table 2, the U_ID 2 quasi-identifier information becomes the background knowledge for an intruder to re-identify U_ID 3 and leads

a path for several attacks as listed in Table 3. Due to the implementation of the 1:1 dataset privacy model on the 1:M dataset, an intruder can gain knowledge from the published data, which causes various correlation attacks such as background knowledge(*bk*) attack, Multiple Sensitive Attribute correlation attack(M*SAcorr*), Quasi-identifier correlation attack(*QIcorr*), Membership correlation attack*(Mcorr)* and Non-Membership correlation attack(*NMcorr*).

3.2 Challenge 2 (Individual Condition Fingerprint Array identification)

In the 1:M dataset, each individual will have multiple records with multiple sensitive attribute(MSA) values and common quasi-identifier values. To anonymize 1:M dataset(challenge 1), the MSA of the individuals are grouped. The grouped MSA with different values alone form an Individual Condition Fingerprint Array identification [ICFA]. The intruder can use this ICFA to re-identify the person with all the sensitive values. It cannot be assured that all individuals ICFA in the dataset will form a unique bucket. A technique, *frequency*-slicing *(f*-slicing*),* has been introduced to deal with different fingerprint array lengths. As per our knowledge, though existing systems have dealt with sensitive attribute fingerprint buckets, they have not adopted any technique that handles the size of the fingerprint array to protect from high utility loss and privacy breach. So, the adversaries can easily blow the Individual Condition Fingerprint Array to get the individual record. The above two challenges on the 1:M dataset make privacy-preserving data publishing very complex with optimal balance between utility and privacy. Achieving high privacy with less information loss in the 1:M dataset is always a challenge and this has been addressed in the paper.

4. 1:M multiple sensitive attribute attacks with corresponding scenarios

The proposed model *f*-slip anonymizes the 1:M dataset with multiple sensitive attributes and guarantees the intensified privacy with minimum loss of information. The proposed approach has been evaluated against various correlation attacks listed in Table 3 by implementing it in the real-world dataset. Five privacy breach cases have been discussed over 2-anonymity published data in Table 2 and explained each case. The implementation of 1:1 dataset privacy models on the 1:M dataset could not resist the following attacks: *bk*, M*SAcorr, QIcorr, Mcorr* and *NMcorr.* The explanation of different cases of the above correlation attacks is as follows.

Table 3 Multi-record multiple sensitive attributes attacks (MMSA)

Multi-record multiple sensitive attributes attacks(MMSA)	**Description**	**Cases**
Background Knowledge *(bk)* attack	An intruder can accomplish *bk* attack if he has strong background knowledge about the individual.	1
Multiple Sensitive attribute correlation attack(M*SAcorr*)	An intruder can accomplish MSAcorr attack; if he knows the value of one sensitive attribute of an individual, he can identify the remaining sensitive attribute values with the help of quasi-identifier and *bk*.	1&2
Quasi-identifier correlation attack(*QIcorr*)	An intruder can accomplish QIcorr attack if he has evident information of QI values and *bk* about the individual.	1&3
Non-membership correlation attack(*NMcorr*)	An intruder can perform *N*Mcorr attack, if he can successfully find the non-existence of an individual in the 1:M dataset.	1,3&4
Membership correlation attack(*Mcorr*)	An intruder can perform Mcorr attack if he can successfully find the existence of an individual in the 1:M dataset.	1,2,3&4

Case 1: An intruder can infer the complete details of the individual's sensitive attributes if he possesses strong background knowledge about the individual. If an intruder knows the basic information of the person, (i.e.) Lisa, is a female, her age is below 20, from zip code 60,000 also possess strong background knowledge such as Lisa has completed her bachelors and earn a decent amount of salary then, he can easily infer the complete information of Lisa which leads to Background Knowledge *(bk)* attack.

Case 2: Multiple sensitive correlation attribute attacks (M*SAcorr*) can occur with the help of background knowledge (*bk*). For example, if an intruder knows that Alan has not studied much and earns less with a high poverty line. The intruder can easily infer Alan's salary and also his disease codes are <V22, V90> from Table 2. Just by knowing one sensitive attribute value, the values of other sensitive attributes can be inferred.

Case 3: If an intruder has the background knowledge about the quasi-identifier values of an individual, he can correlate the QI values of an individual with the values of the sensitive attributes to perform a Quasi-identifier correlation attack (*QIcorr*). Suppose an intruder can correlate or map the sensitive attributes information with the assistance of background knowledge (*bk*) and QI attributes such as age, zip code, sex. In that case, he can infer the sensitive values with high confidence.

Case 4: If an intruder has background knowledge that Dalia is not poor and didn't complete any degree courses, he can find out the QI values of an individual Dalia from Table 2. With the above information and QI values, the intruder can easily infer that Dalia belongs to 2 and 3 equivalence classes. So, Table 2 fails to guarantee privacy and leads to a Non-membership correlation attack (*NMcorr*).

Case 5: If the intruder can infer the existence of an individual along with the complete information, then a membership correlation attack happens. If the intruder knows the QI values of Helen (i.e.) 30, F, 68,000 and possess background knowledge such as Helen's education is higher and poverty is very low, the intruder can easily infer Helen falls in the equivalence classes 3&4 and comes to a conclusion, which record belongs to Helen, with the help of sensitive attribute values.

5. Contribution

During 1:M dataset publishing, the patients' records need to be protected with less information loss. The anonymization methods and models that were implemented earlier to balance the privacy and utility were limited with various factors such as dimensionality, techniques, methodology, etc. The published 1:M dataset becomes ineffective when the quasi-identifier and sensitive attributes are generalized in larger intervals. Moreover, anatomization was performed without considering the correlation between attributes which could lead to the breaking of linking relationship with complete loss of information. The significant contribution of the work is as follows.

1. A thorough study has been done on existing privacy-preserving techniques and models of the 1:M dataset with multiple sensitive attributes to balance privacy and utility.
2. An Anatomization (def.1) is performed based on the correlation between the attributes, which significantly outpaces the breaking of linking relationship concerning privacy and utility of the dataset.
3. After anatomization, Individual Condition Fingerprint Array identification [ICFA] (def.2) is framed and an anonymization method "*f*-slicing" (def.5) has been implemented on sensitive attributes of the 1: M dataset. The "*f*-slicing" can perform based on the frequency of occurrence of ICFA.
4. Based on "*f*-slicing," a privacy model, "*frequency*-slicing with intensified privacy *(f*-slip)" has been proposed to achieve less information loss with intensified privacy.
5. With the experimentations performed on the proposed model, it is proved that "*f*-slip" successfully accomplished intensified privacy with minimum loss of information.

6. Preliminaries

Let R_T be the relational table. R_T encloses explicit identifiers EI= $\{ei_1,ei_2,ei3....ei_n\}$, quasi-identifiers QI=$\{qi_1,qi_2,qi_3...qi_s\}$, and sensitive attributes SA= $\{sa_1,sa_2,sa_3....sa_h\}$.

Definition 1: Anatomy (Xiaokui and Yufei 2006)

Anatomy is a method used to partition the original table into various sub tables. It disconnects the correlation between the quasi-identifiers and sensitive attributes. The main concept behind anatomy is to partition the table, apply different techniques on the quasi-identifier table (QI_T) and sensitive attribute table (SA_T'), and join them together. Both QI_T and SA_T' are assigned with a unique id for reference; however, it will be removed while publishing.

The QI_T has a representation of

$$QI_T = \{Unique\_ID, qi_1, qi_2, qi_3...qi_s\}$$

For each tuple $tp \in R_T$, QI_T has all tuple in the form of

$$(tp[1],tp[2],tp[3],.....,tp[d], Unique\_ID)$$

The SA_T' has a representation of

$$SA_T' = \{Unique\_ID, sa_1', sa_2', sa_3'....sa_h'\}$$

The SA_T' of multi-record (an individual having multiple records) with multiple sensitive attributes has a representation of

SA_T= {Unique_ID, sa_{11}'Usa_{12}'Usa_{13}', sa_{21}'Usa_{22}'Usa_{23}', sa_{31}'Usa_{32}'Usa_{33}'….sa_{h1}'Usa_{h2}'Usa_{h3}'}

Where sa_{11}'Usa_{12}'Usa_{13}' is an aggregation of multiple sensitive values of an individual.

The QI_T is also further divided into sub-tables based on the categorical and numerical values in our approach.

$QI_{T(}$cat) = {Unique_ID, qi_{cat1}, qi_{cat2}, qi_{cat3}…qi_{cats}}
QI_T(num) = {Unique_ID, qi_{num1}, qi_{num2}, qi_{num3}…qi_{nums}}

Definition 2: Individual Condition Fingerprint Array (ICFA)

In a multi-record dataset with multiple sensitive attribute R_T, each individual has multiple values for the sensitive attribute. The various values of sensitive attributes of an individual are written as <MSA_1, MSA_2, $MSA_{3....>}$ are grouped to form Individual Condition Fingerprint Array [ICFA ∈ <MSA_1, MSA_2, $MSA_{3....}$>].

Definition 3: Equivalence Class
For a multi-record dataset $R_{T,}$ the tuples with the same quasi-identifier values in R_T form the equivalence class.

Definition 4: k-anonymity
The published data is said to satisfy k-anonymity if and only if the information of each individual in the published data cannot be distinguished from at least k-1 individuals whose information also appears in the published table.

Definition 5: *f-slicing*
It is an approach to invariably partition high dimensional data based on the similarity between multiple attributes in the dataset using a mode frequency value *f* and forming equivalence classes of size *'f'* containing the similarly grouped attribute tuples.

7. *frequency*- slicing with intensified privacy *(f*-slip*)* Model

After analyzing various existing models, a model *f*-slip has been proposed with an algorithm and architecture framework. It is observed that lots of works have not been carried out in the 1:M dataset with multiple sensitive attributes. It has been proved that the proposed model resists various attacks and those are explained clearly in the experimental evaluation section. The goal of *the f*-slip model is to ensure less information loss with intensified privacy during the privacy preserved data publishing of 1: M with multiple sensitive attributes. The proposed model performs the following steps i) finding the correlation between sensitive attributes ii) pre-processing and aggregation of multi-record to a single record iii) anatomizing both QI and SA iii) implementing k-anonymity on QI iv) *f*-slicing on ICFA and v) Merging of QI and SA.

7.1 Correlation of Sensitive Attributes

In the *f-slip* model, the first step is to find the correlation between the sensitive attributes. The purpose of calculating the correlation between the sensitive attributes is to measure the dependency among the different sensitive attributes. If the anatomization is just performed for splitting the tables into two or three without computing the correlation, it leads to breaking of linking relationship. Suppose the table is divided into multiple tables without computing the correlation between the sensitive attributes. In that case, there are lots of chances that unrelated attributes may be grouped together in the same table, which leads to much information loss. In Table 1, there are four sensitive attributes such as salary, poverty, education and disease code. The poverty, education and disease code are categorical and salary is numerical. Two correlation metrics are used to find the correlation between categorical and numerical.
Cramer's V is a metric used to measure the correlation between the categorical sensitive attributes.

$$CV=\sqrt{\frac{chi^2}{TS*\min(nc-1,nr-1)}} \quad (1)$$

Where chi^2 = chi-square statistics, TS=total sample size, nr=number of rows, nc= number of columns.

One-Way ANOVA is a metric used to measure the correlation between numerical and categorical attributes.

$$\text{OWA}= \frac{MNSQ_B}{MNSQ_W} \tag{2}$$

Where $MNSQ_B$ = mean square between samples, $MNSQ_W$=mean square within sample.

$$\text{MNSQ}_B= \frac{SSS_B}{(ng-1)} \tag{3}$$

$$\text{MNSQ}_W= \frac{SSS_W}{(n_{tob}-ng)} \tag{4}$$

Where SSS_B = Sum of square between sample, SSS_W= Sum of square within sample, ng= number of groups, n_{tob}=total number of observations.

$$\text{SSS}_B= \sum_{ng} n_{ng}(\bar{y}_{ng}-\bar{y})^2 \tag{5}$$

$$\text{SSS}_W= \sum_{x,ng}(y_{x,ng}-\bar{y}_{ng})^2 \tag{6}$$

The correlation between sensitive attribute salary, poverty, education and disease code is calculated.

Table 3 Correlation between the sensitive attribute (SA_T').

		S_1	**S_2**	**S_3**	**S_4**
		Poverty	**Education**	**Disease code**	**Salary**
S_1	**Poverty**	1.000000	0.219627	0.105309	84572.810000
S_2	**Education**	0.219627	1.000000	0.124933	78364.120000
S_3	**Disease code**	0.105309	0.124933	1.000000	18.460000
S_4	**Salary**	84572.810000	78364.120000	18.460000	1.000000

As per the correlation metrics, we have anatomized the sensitive table (SA_T') into two subtables SA_{T1} and SA_{T2}, each with highly correlated attributes. As per Table 3, poverty and salary are highly correlated, so SA_{T1} is formed with poverty and salary and SA_{T2} with education and disease code.

7.2 Pre-processing and aggregation of multi-records.

In the 1:M dataset, the individual records are distributed. Since the records of an individual are widely distributed, we could not form an Individual Condition Fingerprint Array (ICFA) (definition 2). Several records of an individual are aggregated to single record R^{T*} as shown in Table 4. The different sensitive attribute values of an individual are aggregated to form an Individual Condition Fingerprint Array (ICFA). The intruder might use ICFA along with the quasi-identifier values to find the complete information of an individual. For example, if an intruder knows that Tom quasi-identifier values as age are 32, gender is M and zip code is 56,000; also he has completed his masters and earn a good salary with poverty line low, the intruder can easily infer that the last three records of table 2 belong to Tom and can find the complete details of Tom.

Table 4 Aggregated Table R^{T*}.

	Explicit Identifier		**Quasi-identifier**			**Sensitive Attributes**			
	Unique_ ID	**Name**	**Age**	**Gender**	**Zip code**	**Salary**	**Poverty**	**Education**	**Disease code**
rt_1	1	Lisa	19	F	60,000	20,000	medium	Bachelors	<401>
rt_2, rt_3	2	Alan	13	M	55,000	10,201	high	9th	<V22, V90>
rt_4, rt_5	3	Dalia	17	F	70,000	17,258	medium	12th	<724, 408>
rt_6, rt_7	4	Helen	30	F	68,000	25,364	low	Master	<276, 402>
rt_8	5	Tony	16	M	75,000	19,223	medium	11th	<492>
rt_9, rt_{10},rt_{11}	6	Tom	32	M	56,000	21,012	low	Master	<272, 404, 490>

As per the sample dataset, poverty, salary and education do not have multiple values for each individual record. Hence, the ICFA is formed only for the disease code, which has various values. Most of the existing systems have not given importance to preserving ICFA, which leads to privacy breaches. As the sensitive attributes in the 1: M dataset are diversified, the length of the Individual Condition Fingerprint Array for each individual will be unequal. After aggregating the records, the duplicate records are deleted as of pre-processing steps and missed values are filled with the average values of the attributes. In Table 4, the ICFA bucket of Tom is <272, 404, 490>. As the sensitive attributes salary, poverty and education do not have multiple values, and they are not included in the ICFA bucket.

7.3 Anatomization of QI, SA and implementation of k-anonymity on QI_T.

The original 1:M dataset is partitioned into two sub-tables i) Quasi-identifier table (QI_T) and ii) Sensitive attribute table (SA_T'). The quasi-identifier table is further partitioned into two sub-tables, one with categorical attributes (QI_T(cat))and another table with numerical attribute (QI_T(num)). The splitting of the quasi-identifier table is to minimize the utility loss, as we are not generalizing the categorical quasi-identifiers. The numerical attributes are replaced with the mean of the equivalence class (def. 3). For example, if the values of the age in the equivalence class are 33, 35 and 40, then the age values are replaced with 36, which is the mean value. The mean value of an equivalence class is calculated as shown in equ.7.

$$\text{Mean} = \frac{qi_{11} + qi_{12} + qi_{13}}{n} \tag{7}$$

Where qi_{11}, qi_{12}, qi_{13} are the quasi-identifier values of age in an equivalence class and n is the number of values in an equivalence class.

$$\text{Mean (Age } (E_{qc})) = \frac{33+35+40}{3} = 36 \tag{8}$$

Where Mean (Age (E_{qc})) is the mean value of age in an equivalence class, similarly, the mean value is calculated for each equivalence class.

$$\text{Mean (Zip code } (E_{qc})) = \frac{60000+55000+56000}{3} = 570000 \tag{9}$$

Where Mean (Zip code (E_{qc})) is the mean value of zip code in an equivalence class.
K-anonymity (def. 4) model is implemented on both QI_T(num) and QI_T(cat) separately and a unique id is generated for all the tuples, as shown in Tables 5&6. Finally, the QI_T(num) and QI_T(cat) are merged to form the QI_T(final) table. As per our knowledge, we are the first to anatomize the quasi-identifier table in the 1:M dataset to reduce the loss of information.

Table 5 2-anonymity QI_T(num)

Unique_ ID	Age	Zip code
1	16	57,500
2	16	57,500
3	23	69,000
4	23	69,000
5	24	65,500
6	24	65,500

Table 6 2-anonymity QI_T(cat)

Unique_ ID	Gender
1	F
2	M
3	F
4	F
5	M
6	M

7.4 *f*-slicing and merging of QI and SA.

The sensitive attribute table is partitioned according to the correlation of sensitive attributes. As per correlation [Education and Disease Code $\in SA_{T1}$] [Salary and Poverty $\in SA_{T2}$], as shown in Table 7a &8a. The categorical values of sensitive attributes are anonymized using a bottom-up approach and numerical SA values are anonymized by replacing the actual value with the mean values of the equivalence class. Education and poverty are generalized, as shown in Figures 2&3. The disease codes are generalized and the sample is shown in Figure 1. The related diseases for the disease codes are mapped using the ICD9X dictionary. As the ICD9X dictionary cannot be imported in python 3 , a new dictionary has been framed according to the dataset with the help of the ICD9X dictionary. Since there are several disease codes available, the values are generalized and replaced in SA_T'.

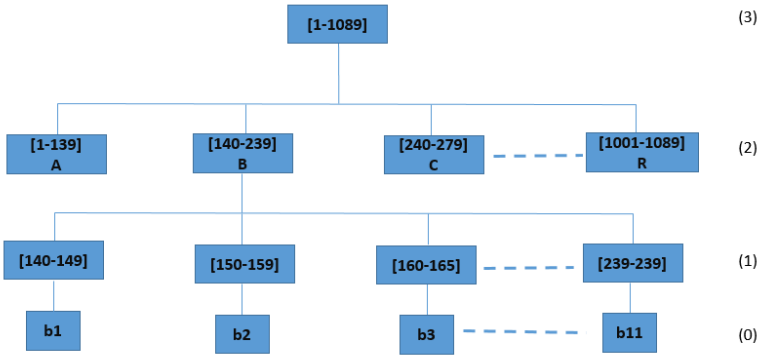


Fig. 1 Bottom-up approach generalization for disease code attribute .

The ICFA buckets of disease code are arranged alphabetically and the number of occurrences of each bucket is stored to calculate the frequency of occurrence of each ICFA bucket to set the value of *f*. The SA_{T1} is sorted with respect to education and disease code and equivalence classes are formed according to the frequency calculated. A grouped (GID) is allotted to all the equivalence classes created with similar records. As per SA_{T1}, the frequency *f*=2, since the disease codes are categorical, the maximum similarities between the fingerprint buckets in an equivalence class are taken. Few disease codes have the V alphabet in them; therefore, V's value is taken as 4; thus, the alan disease codes have been converted to 422 and 490. For $SA_{T2,}$ the value of *f* = 3, since the salary is numerical, the average of the equivalence class has been taken as shown in table 7b & 8b. A group id (GID) is also created for the equivalence classes formed according to the frequency calculated. The unique_id is created just for reference because the records get shuffled during anonymization.

Table 7 Sensitive attribute table of Salary and Poverty

Unique _ID	Salary	Poverty
2	10,201	high
3	17,258	medium
5	19,223	medium
1	20,000	medium
6	21,012	low
4	25,364	low

(a) Sensitive attribute SA_{T1}*

Unique_ID	Salary	Poverty	GID
2	15,570	high	1
3	15,570	medium	1
5	15,570	medium	1
1	22,125	medium	2
6	22,125	low	2
4	22,125	low	2

(b) Sensitive attribute SA_{T1}** f= 3 (salary is numerical so, an average of the equivalence classes has been taken)

Table 8: Sensitive attribute table of Education and Disease code

Unique _ID	Education	Disease code
2	9th	422(C), 490(D)
5	11th	492(D)
1	Bachelors	401(D)
3	12th	724(G), 408(D)
4	Master	276(B), 402(D)
6	Master	272(B),404(D),490(D)

(a) Sensitive attribute SA_{T2}*

Unique_ID	Education	Disease code	GID
2	9th	C,D	1
5	11th	C,D	1
1	Bachelors	G,D	2
3	12th	G,D	2
4	Master	B,D,D	3
6	Master	B,D,D	3

(b) Sensitive attribute SA_{T2}** f =2 (since disease codes are categorical so, the maximum similarities between fingerprint buckets in an equivalence class are taken)

The novel approach in the *f*-slip model is the slicing of records according to the frequency of occurrences of sensitive attribute values in each sub-table. In the existing works, records in the anatomized table are not grouped (sliced) together according to the frequency of occurrences and the partitioned tables are sliced with a fixed variable. In contrast, the *f*-slip model anonymizes the partitioned table with dynamic variable *f* named *"f-slicing*." After the anonymization of SA_{T1} and SA_{T2}, both the tables are merged along with the GID. The complete framework of *the f*-slip model is depicted in figure 4. The anonymized tables that need to be disclosed with privacy preservation are shown in table 9. The privacy-preserving data publishing of the *f*-slip model needs to disclose the tables in three partitions a. QI_T_final, b. Sensitive attribute SA_{T1} and c. Sensitive attribute SA_{T2}. As the sensitive attributes are anonymized using *f*-slicing, the mapping of individual details through Table 9a, 9b, 9c is very difficult.

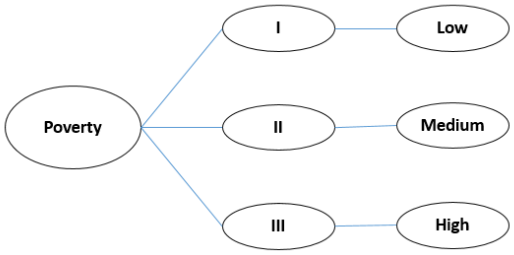


Fig. 2 Generalization of attribute poverty.

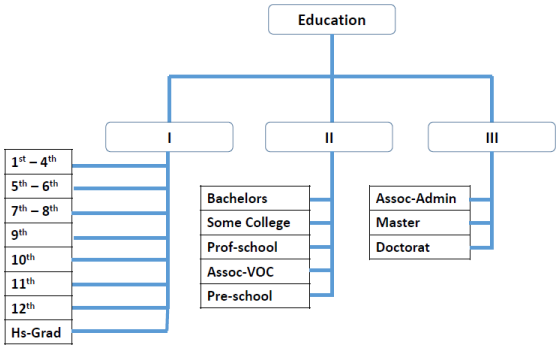


Fig .3 Generalization of attribute education.

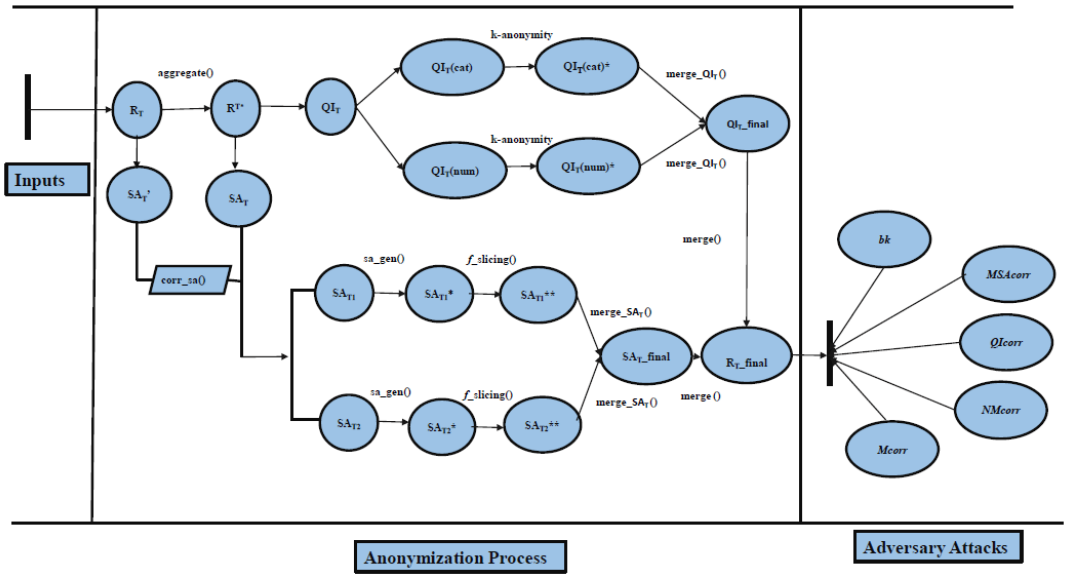


Fig. 4 Framework of *the f*-slip model.

Table 9 RT_A anonymized data.

Age	Zip code	Gender
16	57,500	F
16	57,500	M
23	69,000	F
23	69,000	F
24	65,500	M
24	65,500	M

(a) QI_T_final

Salary	Poverty	GID
15,570	high	1
15,570	medium	1
15,570	medium	1
22,125	medium	2
22,125	low	2
22,125	low	2

(b) Sensitive attribute SA_{T1}

Education	Disease code	GID
9th	C,D	1
11th	C,D	1
Bachelors	G,D	2
12th	G,D	2
Master	B,D,D	3
Master	B,D,D	3

(c) Sensitive attribute SA_{T2}

8. *f*-slip Algorithm

The main objective of the *f*-slip algorithm is to provide a balance between privacy and utility. As per our knowledge, the existing systems have implemented anatomization only in the sensitive attribute table for 1:M datasets, whereas *f*-slip performs anatomization in both quasi-identifier and sensitive attribute table. Anatomizing the SA table alone persists in various privacy breaches.

Algorithm 1: *f*-Slip
Input: R_T, k
Output: R_T_final
(1) *f*-Slip (R_T) :
// 1:1 Converted R_T
(2) R_T*<- *aggregate*(R_T) ;
// Splits R_T into QI_T, SA_T
(3) anatomize(R_T*)
// Spliting of R_T* into QI_T , SA_T
// Anonymization of QI table
(4) QI_T_final <-*anon_QI*(QI_T ,k) ;
// correlation between Sensitive attributes
(5) D <- *corr_sa*(SA_T')
// Anonymization of SA table
(6) SA_T_final <- *anon_SA* (SA_T , D) ;
// Merging of QI and SA table
(7) R_T_final <- *merge* (QI_T_final, SA_T_final);
(8) return R_T_final ;

Algorithm 2: anon_QI(QI_T, k)
Input: QI_T, k
Output: QI_T_final
// Anonymizing the quasi-identifier table
(1) anon_QI (QI_T, k)
// Splits QI_T into QI_T (num), QI_T (cat)
(2) anatomize(QI_T)
// Applying k-anonymity QI_T (num), QI_T (cat)
(3) QI_T (num)* <- *k_anonymity* (QI_T (num), k) ;
(4) QI_T (cat)* <- *k_anonymity* (QI_T (cat), k) ;
// Merging the k-anonymized QI_T (num), QI_T (cat)
(5) QI_T_final <- *merge*(QI_T (num)*, QI_T (cat)*) ;
(6) return QI_T_final;

The process steps of the proposed model *f*-slip are split into four algorithms (1,2,3 and 4) and outlined for understanding purpose. In Algorithm 1, the relational table R_T is passed as input in line 1. The multi-records of an individual are aggregated in line 2. The aggregated table R_T * is anatomized into the quasi-identifier table and sensitive attribute table in line 3. Further, the quasi-identifier table is k-anonymized by passing the value of the k parameter in line 4. The correlation between the sensitive attributes is computed and anonymized based on the correlation in lines 5& 6. Finally, the anonymized quasi-identifier and sensitive attribute table are merged and returned in lines 7&8. The implementation of k-anonymity to anonymize the quasi-identifier table is depicted in algorithm 2. The quasi-identifier table and k parameter are passed as an input parameter in line 1. The QI_T is anatomized into QI_T (num) and QI_T (cat) in line 2. Both QI_T (num), QI_T (cat) are anonymized separately by implementing k-anonymity and merged together from lines 3-6.

In algorithm 3, the correlation between the sensitive attribute is calculated. In line 1, the partitioned table SA_T' is passed as an input to the correlation function. To find the correlation between sensitive attributes, the attributes are categorized into both categorical and numerical in line 2. Cramer's v is applied on categorical attribute in line 3 and one-way ANOVA is implemented on both categorical and numerical attributes and stores in variable D from lines 4-8. The anonymization of the sensitive attribute table is outlined in algorithm 4. The aggregated table SA_T and correlation table D are passed as an input parameter in line 1. Based on the correlation, the SA table is anatomized and generalized from lines 2-4. The generalized tables are sorted to find the *f* parameter in lines 5&6. The process of setting the frequency mode for both categorical and numerical attributes is explained from lines 7-15. The *f*-slicing is performed on both SA sub tables based on the frequency of occurrences of sensitive attributes

and group-id is allotted from lines 16-23. Finally, the merging of f-sliced sub tables is performed in line 24. The functions used in the algorithms are described in Table 10.

9. Experimental evaluation and result analysis

This section presents the experimental evaluation and results from the analysis of the proposed model on the real-world dataset in terms of time execution and data utility. The result analysis of the proposed model is compared with the existing 1:M privacy models through graphs. The implementation of *the f*-slip **model** is carried out in Python 3 and experimental evaluation and result analysis are executed in the windows 10 operating system with 8 GB memory. The implementation and evaluation are performed on real-world dataset Informs. In Informs dataset, {age, racex, sex, marry} are taken as QI attributes and { income, poverty, education, condition_code} as sensitive attributes.

Algorithm 3: corr_sa (SA_T')

Input: SA_T'
Output: D
// Finding the correlation between the sensitive attributes
(1) corr_sa (SA_T') :
// Anatomizing the sensitive attributes based on the correlation
(2) *categorise*(SA_T')
// Splits SA_T' into SA_T' (num), SA_T' (cat)
// correlation between the categorical sensitive attributes
(3) c <- *cramers_v* (SA_T'(cat)) ;
// correlation between the numerical and categorical sensitive attributes
(4) for i,j in SA_T' (num), SA_T' (cat):
(5) o <- *owa*(i,j) ; // One Way Anova (f-test)
(6) end for
(7) D <- c,o
(8) return D ;

Algorithm 4: anon_SA ()

Input: SA_T , D
Output: SA_T_final
// Anonymization of SA
(1) anon_SA (SA_T, D) :
// Anatomizing the SA_T by using the correlation value D
(2) SA_{T1} , SA_{T2} <- *anatomize*(SA_T,D)
// Generalizing the attribute value in SA_{T1} & SA_{T2}
(3) SA_{T1}* <- *sa_gen*(SA_{T1})
(4) SA_{T2}* <- *sa_gen*(SA_{T2})
// Sorting the values of SA_{T1}* & SA_{T2}*
(5) *sort* (SA_{T1}*)
(6) *sort* (SA_{T2}*)
// Setting frequency mode for both categorical and numerical.
(7) if(SAT1 = categorical)
(8) f1 = med(fi(SAT1*)– {1})
(9) else if(SAT1 = numerical)
(10) f1 = mean(fi(SAT1*)-{1})
(11) if(SAT2 = categorical)
(12) f2 = med(fi(SAT2*)– {1})
(13) else if(SAT2 = numerical)
(14) f1 = mean(fi(SAT2*)-{1})
(15) end if
//*f*-slicing
(16) GID=0
(17) for i in SA_{T1}*
(18) SA_{T1}** <- SA_{T1}*[i:i+f1] -> GID+1
(19) GID2=0
(20) end for
(21) for i in SA_{T2}*
(22) SA_{T2}** <- SA_{T2}*[i:i+f2] -> GID2+1
(23) end for
(24) SA_T_final <- *merge*(SA_{T1}**,SA_{T2}**)
(25) return SA_T_final ;

Table 10 Summary of all functions used in the algorithms.

Functions	Description
aggregate()	Aggregating the multiple records of an individual into a single record.
anatomise()	Anatomizing the original table into QI_T and SA_T.'
corr_sa ( )	Computing correlation between the sensitive attributes.
anon()	Anonymizing the partitioned tables.
merge()	Merging the anonymized table.
sa_gen ()	Generalized the sensitive attribute tables.
*f*_slicing()	Performing the *f*_slicing on sensitive attribute sub tables.

During the pre-processing phase, the duplicate records are removed and the missing values are replaced with the average values of the attributes. There are 2,44,321 records in Informs datasets. After aggregation and pre-

processing, the total number of records in the dataset is 42,999. Since the dataset is composed of both categorical and numerical, the utility is measured for both categorical and numerical separately. The primary challenge faced in the *f*-slip is making the dictionary for condition_code and fixing of 'f' value for SA_{T1}* and SA_{T2}*.

9.1 Utility Loss

The utility of anonymized data is measured for both categorical and numerical separately. The metric Normalized Certainty Penalty (NCP) (Qiyuan et al. 2017) is used to measure the information loss for the categorical attributes in QI_T and SA_T. The utility metric measure for the numeric data is Numeric Information Loss (NIL) which is framed newly by customizing the iloss metric (Xinning and Zhiping 2020). NIL is used to measure the information loss for QI_T and SA_T numerical attributes.

$$\text{NCP}(e) = \frac{|e|}{|n|} \tag{10}$$

Where $|e|$ is the nodes covered by the generalized value and $|n|$ is the total number of nodes.

The output of NCP ranges from 0 to 1. The value 0 means no information loss and as the output value of NCP increases, the information loss also increases. The NCP output is directly proportional to the information loss. The NCP for each node is calculated as per equ.10. For example, the attribute poverty has five values and those five values are generalized to 4 nodes and the NCP calculation for each node is shown in fig.5.

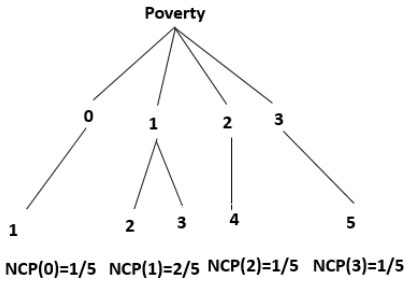


Fig. 5 Generalization hierarchy for attribute poverty

For detailed explanation of NCP measure, the information loss in condition_code is calculated as below:

Informationloss (condition_code) = No.of unique values of 0 * NCP (0)

+ No.of unique values of 1 * NCP (1)

+ No.of unique values of 2 * NCP (2)

+ No.of unique values of 3 * NCP (3)

The total number of unique condition_codes(U_{cc}) are

U_{cc} = {001, 002 … 201 ….. 927, 928 ….1038} =1038.

Figure 6a depicts the generalization hierarchy for condition_code and fig. 6b shows the generalization of node 927. Though the dataset has 1-1038 condition codes, it is not continuous; many condition_code values in between are missing. In fig. 6, we can observe in node B, the range of values is 140-239, so the value present in the range should be 99 codes; however, the values present are 94 and five values are missing in between. Therefore, the range of values cannot be fixed as they are not consistent for each node. NCP for node Q is calculated as per equ.11.

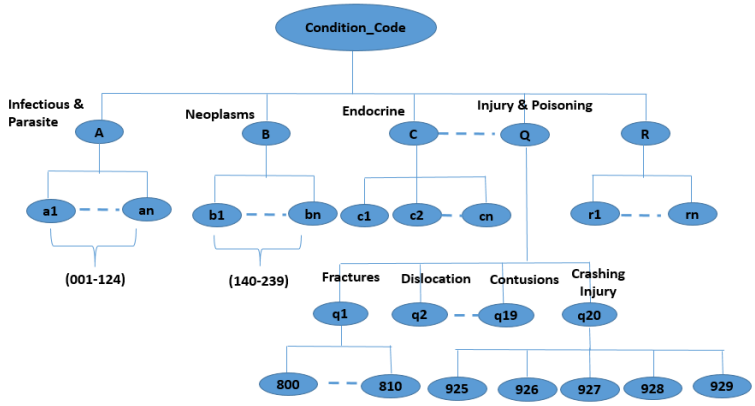


Fig. 6a Generalization hierarchy for disease code

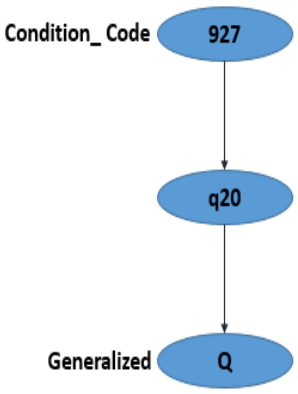


Fig. 6b Generalization of node 927

The total diseases in each node are A: 124; B: 94; …. Q: 202; R: 89.

$$\text{NCP (Q)} = \frac{No.of\ nodes enclosed by\ Q}{Total\ Nodes} = \frac{202}{1038} = 0.194 \quad (11)$$

The equ.11 has calculated the NCP for node Q and resulted in less information loss. The information loss for three conditions h, h and b are calculated as per equ.12.

$$\text{NCP (HHB)} = \frac{\left(\frac{51}{1038} + \frac{51}{1038} + \frac{94}{1038}\right)}{3} = 0.062 \quad (12)$$

The average sum of information loss for three conditions h, h and b are 0.062, which is very less. The average information loss of categorical attributes is 14.43%.

The information loss of all records in relational table R_T for numerical attributes is calculated using a metric NIL:

$$\text{NIL } (R_T, A_{na}) = \frac{\sum_{i}^{n} | old\ R_T(A_{na})_i - new\ R_T(A_{na})_i |}{Unique(R_T)}$$

Where R_T is the relational table, old (R_T) is the original data before anonymization, new (R_T) is the new data after anonymization, unique (R_T) is the unique values in the numerical attribute and n is the number of tuples in R_T. The total information loss in the numerical attribute is 7.97% as per the metric NIL.

NCP and NIL have been used to measure the information loss on categorical and numerical attributes, respectively. For dataset Informs, we have analyzed the QI-information loss and SA-information loss with varying sizes of the dataset, as shown in Figure 7. To evaluate whether the data utility is subtle as the size of the dataset (n) increases. A series of data subsets are randomly selected from the whole dataset in intervals of 5000. The *f*-slip algorithm has been run for different sizes of the dataset. It is clearly shown that RMR has a high information loss for different sizes of the dataset and 1:M Mondrian has 30% of information loss for the 5k sample dataset. As the size of the dataset increases, the information loss is less and consistent. The *f*-slip model results in very less information loss and the loss are consistent as the size of the dataset increases for both QI_T and SA_T.

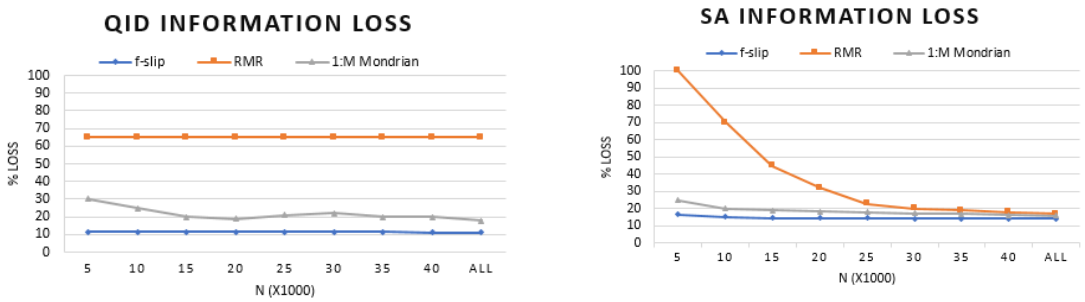
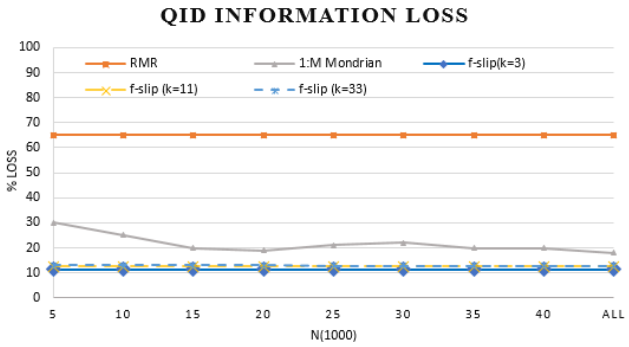


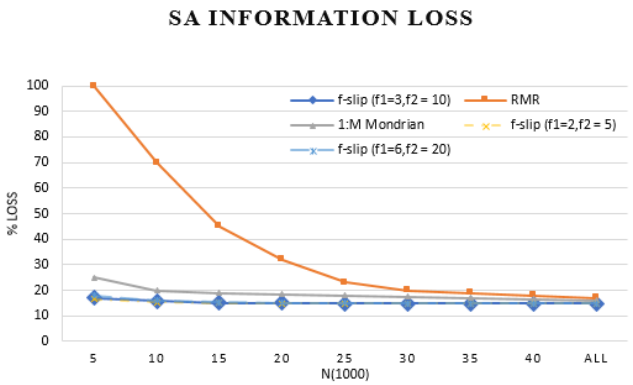
Fig. 7 Information loss while varying data size(n).

The existing model RMR results in 66% of QI-NCP by setting the value δ=0.66. RMR has much information loss in quasi-identifier as it concentrates more on preserving the sensitive attributes. As shown in fig. 7, RMR has a high utility loss compared to 1:M Mondrian and *f*-slip model. The information loss in *f*-slip is approximately 9% and 4% lesser than 1:M Mondrian in QI and SA, respectively.

The QI-information loss has been computed with varying parameter k and depicted in figure:8a. As shown in figure 8a, it is evidently shown that the RMR has high utility loss in all cases. As the proposed model does not have a k parameter in SA, the information loss in QI is alone computed by varying parameter k. In 1:M Mondrian, both QI-NCP and SA-NCP are sensitive to parameter k; thus, there is variation in QI-utility loss. The *f*-slip model has not implemented generalization in the QI table so, there is less information loss and it is consistent though the k parameter increases. The information loss in the sensitive attribute is calculated by varying parameter *f*, as shown in fig 8b. The RMR and 1:M Mondrian do not have *f* parameters, so the information loss percentage based on the number of records has been plotted in fig. 8b. The sensitive attribute sub tables have been executed by fixing different *frequency* values for different records. As shown in fig.8b though the value of *f* varies, the information loss is very less and consistent. So, the proposed work proves that *f*-slip can work for the higher dimensional dataset with different frequency values.



a. Information loss when varying k parameter



b. Information loss when varying *f* parameter

Fig. 8 Information loss.

9.2 Execution time

The efficiency of the proposed model has been evaluated by the execution time compared with RMR and 1:M Mondrian. The execution time does not involve the pre-processing steps. The proposed model has been executed for a different size (n) of the dataset and compared with RMR and 1:M Mondrian. The captured results are clearly depicted in Fig 9. The *f*-slip is much efficient than RMR. The average execution time of RMR is approximately 1.2 h, the average execution time of 1:M Mondrian is 30 s and the average execution time of *the f*-slip model is 28 s. Though the efficiency of the proposed model is much higher than RMR, *f*-slip and 1:M Mondrian execution times differ in very few seconds. Thus, the average running time of *the f*-slip model and 1:M Mondrian is O(nlogn).

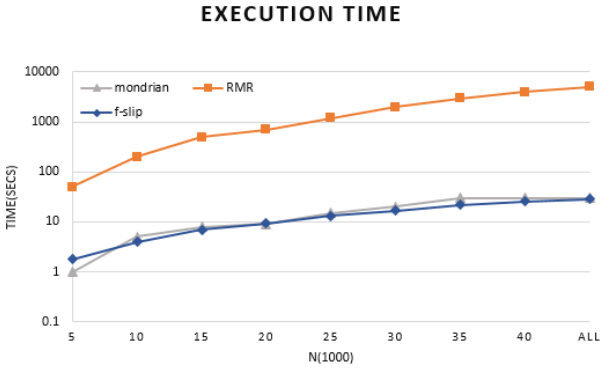


Fig. 9 Execution Time

10 Conclusion and Future Directions

The study presents the work on privacy-preserving data publishing on 1:M datasets. An efficient proposed model named *f*-slip has been proposed to address various attacks such as Background Knowledge (bk) attack, Multiple Sensitive attribute correlation attack(MSAcorr), Quasi-identifier correlation attack(QIcorr), Non-membership correlation attack(NMcorr) and Membership correlation attack(Mcorr). Anatomization is performed to partition the original microdata into two tables' a. quasi-identifier and b. sensitive attribute. After partitioning, k-anonymity has been implemented on the quasi-identifier table. Based on the correlation among the sensitive attributes, the sensitive attribute table is partitioned. An anonymization method, *f*-slicing, was proposed to anonymize the sensitive attributes, whereas the parameter *f* is fixed based on the occurrences of ICFA. The parameter *f* can be fixed dynamically according to the dimensionality of the dataset. Extensive experiments have been performed on a real-world dataset Informs to show that the *f*-slip model performs better than RMR and 1:M Mondrain in terms of efficiency and information loss by varying the size of the datasets. In the study, there are also few limitations in the *f*-slip model. In Informs dataset, the quasi-identifier of all the records of an individual has the same value, whereas it may change in time. E.g. age increases and zip code may also change. Also, in Informs dataset, the only condition_code has multiple values, whereas salary, poverty and education share the same value; thus, ICFA is formed only for condition_code. Extending the work so that all the sensitive attributes might have different values and therefore all the sensitive attributes values can be included in ICFA and the parameter *f* needs to be carefully chosen and fixed.

Declaration

Ethical Approval : This article does not contain any studies with human participants or animals performed by any of the authors.
Funding Details : No Funding is received for this work.
Conflicts of interest/ competing interest: All authors declare that they have no conflict of interest.
Informed Consent : We have not used any content which requires any consent.

Authors' Contribution:
Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing-Original Draft – Jayapradha.J

Conceptualization, Data Curation, Formal Analysis, Investigation, Project Administration, Resources, Supervision, Validation, Writing- Review & Editing - Prakash.M

References

[1] Abdul Majeed., Sungchang Lee.: Anonymization Techniques for privacy preserving data publishing: A Comprehensive Survey. IEEE Access, 9, 8512-8545 (2020). https://doi.org/10.1109/ACCESS.2020.3045700.

[2] Adeel Anju., Naveed Ahmad., Saif U. R. Malik,Samiya Zubair., Basit Shahzad.: An efficient approach for publishing micro data for multiple sensitive attributes. The Journal of Supercomputing. 74, 5127–5155 (2018). https://doi.org/10.1007/s11227-018-2390-x.

[3] Adeel Anjum., Nayma Farooq., Saif Ur Rehman Malik., Mansoor Ahmed., Abid Khan., Moneeb Gohar.: An Effective Privacy Preserving Mechanism for 1: M Microdata with High Utility. Sustainable Cities and Society45. 1-22 (2018). https://doi.org/10.1016/j.scs.2018.11.037

[4] Ashwin Machanavajjhala., Daniel Kifer., Johannes Gehrke., Muthuramakrishnan Venkitasubramaniam.: L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data. 1, 1-52 (2006). https://doi.org/10.1145/1217299.1217302.

[5] Chu Jianxun., Vincent Ekow Arkorful., ZhaoShuliang.: Electronic health records adoption: Do institutional pressures and organizational culture matter. Technology in Society, 65 (2021). https://doi.org/10.1016/j.techsoc.2021.101531.

[6] Ismail Keshta., Ammar Odeh.: Security and privacy of electronic health records: Concerns and challenges. Egyptian Informatics Journal, 1-7 (2020). https://doi.org/10.1016/j.eij.2020.07.003.

[7] Jayapradha. J., Prakash. M, Yenumula Harshavardhan Reddy.: Privacy Preserving Data Publishing for Heterogeneous Multiple Sensitive Attributes with Personalized Privacy and Enhanced Utility. Systematic Reviews Pharmacy 11(9), 1055-1066 (2020). https://doi.org/10.31838/srp.2020.9.151.

[8] Jayapradha. J., Prakash. M.: An efficient privacy-preserving data publishing in health care records with multiple sensitive attributes. Sixth International Conference on Inventive Computation Technologies. 623-629, 2021. https://doi.org/10.1109/ICICT50816.2021.9358639.

[9] Jianneng Cao., Panagiotis Karras.: Publishing micro data with a robust privacy guarantee. International Conference on Very Large Data Bases. 5(11), 1388-1399 (2012).

[10] Jinwen Liang., Zheng Qin., Sheng Xiao., Jixin Zhang., Hui Yin., Keqin Li.: Privacy-preserving range query over multi-source electronic health records in public clouds. Journal of Parallel and Distributed Computing, 135, 127-139 (2020). https://doi.org/10.1016/j.jpdc.2019.08.011.

[11] Junqing Le., Di Zhang., Nankun Mu, Xiaofeng Liao., Fan Yang.: Anonymous privacy preservation based on m-signature and fuzzy processing for real time data release. IEEE Transaction Systems Man, Cybernetics Systems. 50(10), 1–13 (2018). https://doi.org/10.1109/TSMC.2018.2872902.

[12] Khaled El Emam., Fida Kamal Dankar.: Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association, 15, 627-637 (2008). https://doi.org/10.1197/jamia.M2716.

[13] Kristen LeFevre., David J DeWitt., Raghu Ramakrishnan.: Incognito: Efficient full-domain k-anonymity. In Proceedings of the ACM SIGMOD international conference on Management of data, 49-60 (2005).

[14] Lakshmipathi Raju N.V.S., Seetaramanath M.N., Srinivasa Rao P.: An enhanced dynamic KC-slice model for privacy preserving data publishing with multiple sensitive attributes by inducing sensitivity. Journal of King Saud University –Computer and Information Sciences. 1-13 (2018). https://doi.org/10.1016/j.jksuci.2018.09.013.

[15] Latanya Sweeney.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness Knowledge based Systems. 10 (5), 571–588 (2002). https://doi.org/10.1142/S021848850200165X.

[16] Latanya Sweeney.: K-Anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems. 10 (5), 557–570 (2002). https://doi.org/10.1142/S0218488502001648.

[17] Lin Yao., Zhenyu Chen., Xin Wang., Dong Liu., Guowei Wu.: Sensitive Label Privacy Preservation with Anatomization for Data Publishing. IEEE Transactions on Dependable and Secure Computing. 18, 1-14 (2021). https://doi.org/10.1109/TDSC.2019.2919833.

[18] Lin Zhang ., Jie Xuan ., Si Ruoqian., Ruchuan Wang.: An improved algorithm of individuation K-anonymity for multiple sensitive attributes. Wireless Personal Communications 95(3):2003–2020 (2017). https://doi.org/10.1007/s11277-016-3922-4.

[19] Ninghui Li., Tiancheng Li., Suresh Venkatasubramanian.: t-closeness: Privacy beyond k-anonymity and l-diversity. IEEE 23rd International Conference on Data Engineering. IEEE. 106–115 (2007). https://doi.org/10.1109/ICDE.2007.367856.

[20] Orna Fennelly., Caitriona Cunningham., Loretto Grogan., Heather Cronin., Conor O'Shea., Miriam Roche., Fiona Lawlor., Neil O'Hare.: Successfully implementing a national electronic health record: a rapid umbrella review. International Journal of Medical Informatics, 144, 1-17 (2020). https://doi.org/10.1016/j.ijmedinf.2020.104281.

[21] Poulis G., Loukides G., Gkoulalas-Divanis A., Skiadopoulos S.: Anonymizing data with relational and transaction attributes. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. (2013). https://doi.org/10.1007/978-3-642-40994-3_23.
[22] Qiyuan Gong., JunzhouLuo., MingYang., WeiweiNi., Xiao-BaiLi.:Anonymizing 1: M microdata with high utility. Knowledge-Based Systems. 115, 15-26 (2017). https://doi.org/10.1016/j.knosys.2016.10.012.
[23] Rashad Saeed., Azhar Rauf.: Anatomization through Generalization (AG): A Hybrid Privacy- Preserving Approach to Prevent Membership, Identity and Semantic Similarity Disclosure Attacks. International Conference on Computing, Mathematics and Engineering Technologies, 1-7, (2018). https://doi.org/10.1109/ICOMET.2018.8346323
[24] Razaullah Khan., Xiaofeng Tao., Adeel Anjum., Haider Sajjad., Saif ur Rehman Malik., Abid Khan., Fatemeh Amiri.: Privacy Preserving for Multiple Sensitive Attributes against Fingerprint Correlation Attack Satisfying c-Diversity. Wireless Communications and Mobile Computing. 1-18 (2020).
[25] Shyamala Susan V., Christopher T.: Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes. SpringerPlus. 5(1), 964–984 (2016). https://doi.org/10.1186/s40064-016-2490-0.
[26] Stephen L. Meigs., Michael Solomon.: Electronic Health Record Use a Bitter Pill for Many Physicians. Perspectives in Health Information Management. 13 (2016).
[27] Tehsin Kanwal., Adeel Anjuma., Saif U.R. Malik., Haider Sajjada., Abid Khanc., Umar Manzoor., Alia Asheralievae.: A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes. Computers and Security. 105, 1-21(2021). https://doi.org/10.1016/j.cose.2021.102224.
[28] Wang Rong., Yan Zhu., Tung-Shou Chen., and Chin-Chen Chang.: Privacy-Preserving Algorithms for Multiple Sensitive Attributes Satisfying t-Closeness. Journal of Computer Science and Technology, 33(6), 1231–1242 (2018). https://doi.org/10.1007/s11390-018-1884-6
[29] Widodo., Eko Kuswardono Budiardjo., Wahyu Catur Wibowo.: Privacy Preserving Data Publishing with Multiple Sensitive Attributes based on Overlapped Slicing. MDPI information. 10, 1-18 (2019). https://doi.org/10.3390/info10120362
[30] Widodo., Wahyu Catur Wibowo.: A Distributional Model of Sensitive Values on p-Sensitive in Multiple Sensitive Attributes. 2nd International Conference on Informatics and Computational Sciences. 1-5 (2018). https://doi.org/10.1109/ICICOS.2018.8621698.
[31] Xiangwen Liu., Qingqing Xie., Liangmin Wang., Personalized extended (alpha, k)-anonymity model for privacy preserving data publishing. Concurrency and Computation: Practice and Experience. 29(6), 1-18 (2017). https://doi.org/10.1002/cpe.3886.
[32] Xiaokui Xiao., Yufei Tao.: Anatomy: Simple and effective privacy preservation. Proceedings of the 32nd international conference on Very large databases. 139-150. (2006).
[33] Xinning Li., Zhiping Zhou.: A generalization model for multi-record privacy preservation. Journal of Ambient Intelligence and Humanized Computing. 11, 2899–2912 (2020). https://doi.org/10.1007/s12652-019-01430-y.
[34] Yuelei Xiao., Haiqi Li.: Privacy Preserving Data Publishing for Multiple Sensitive Attributes Based on Security Level. MDPI Information. 11, 1-27 (2020). https://doi.org/10.3390/info11030166.
[35] Zhe Xiao., Xiuju Fu., Rick Siow Mong Goh.: Data Privacy-Preserving Automation Architecture for Industrial Data Exchange in Smart Cities. IEEE Transactions on Industrial Informatics, 14(6), 1-11 (2018). https://doi.org/10.1109/TII.2017.2772826.