Privacy Preserving for Multiple Sensitive Attributes against Fingerprint Correlation Attack Satisfying c-Diversity

Razaullah Khan, Xiaofeng Tao, Adeel Anjum, Haider Sajjad, Saif ur Rehman Malik, Abid Khan, and Fatemeh Amiri

1National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing, China
2Department of Computer Science, COMSATS University Islamabad, 45550 Islamabad, Pakistan
3Cybernetica AS Estonia, Tallinn, Estonia
4Department of Computer Engineering, Hamedan University of Technology, Hamedan, Iran

Correspondence should be addressed to Xiaofeng Tao; taoxf@bupt.edu.cn

Received 6 November 2019; Accepted 19 December 2019; Published 28 January 2020

Copyright © 2020 Razaullah Khan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Privacy preserving data publishing (PPDP) refers to the releasing of anonymized data for the purpose of research and analysis. A considerable amount of research work exists for the publication of data, having a single sensitive attribute. The practical scenarios in PPDP with multiple sensitive attributes (MSAs) have not yet attracted much attention of researchers. Although a recently proposed technique (p, k)-Angelization provided a novel solution, in this regard, where one-to-one correspondence between the buckets in the generalized table (GT) and the sensitive table (ST) has been used. However, we have investigated a possibility of privacy leakage through MSA correlation among linkable sensitive buckets and named it as “fingerprint correlation (fcorr) attack.” Mitigating that in this paper, we propose an improved solution “(c, k)-anonymization” algorithm. The proposed solution thwarts the fcorr attack using some privacy measures and improves the one-to-one correspondence to one-to-many correspondence between the buckets in GT and ST which further reduces the privacy risk with increased utility in GT. We have formally modelled and analysed the attack and the proposed solution. Experiments on the real-world datasets prove the outperformance of the proposed solution as compared to its counterpart.

1. Introduction

Data generation and sharing have shown a drastic increase in the ongoing decade. The reason behind is obviously the growing sources of data due to huge research and smart revolution (smart grids, cities, devices, etc.). The utility of the shared/published data is utilized in research and analysis by the data researchers. The research and analysis may involve data mining, statistical data analysis, and other policy makings. In the context of health records, the data owners are the individuals to whom the data belong. The hospital that collects, manipulates, and shares that data is known as the data publisher. The data researchers may be a wide range of stakeholders (e.g., pharmaceuticals, government agencies, and survey organizations). The collected data contain private information (e.g., name, contact number, and social security number), partial identifiers (e.g., age, gender, zipcode, and country), and confidential or sensitive information (e.g., disease) about the data owners. Sharing such sensitive information is a privacy breach and legislatively wrong, if disclosed to unauthorized parties.

To ensure privacy of such information, most of the existing algorithms [1–6] in the literature deal exist with a single sensitive attribute only. However, a dataset may practically have multiple sensitive attributes (MSAs) [7–14]. For example, a hospital may publish data with more than one sensitive attribute, such as disease, symptom, and physician as shown in Table 1. The sensitive nature of healthcare records urges researchers to handle such scenarios and assure that the privacy of an individual may not be breached.
In data publications, along with privacy, data utility is also a major concern so that researchers may perform research and analysis. Therefore, data should be anonymized in such a way that the research analysts may extract useful information. Balancing privacy and utility in privacy preserving data publishing (PPDP) is a NP-hard problem [15–20]. Therefore, the scenario in this paper is more challenging, as we consider the dimensionality in quasi-identifiers (QIs) as well as more than one sensitive attributes, i.e., MSAs.

An adversary or an attacker is a person who tries to breach the data privacy using different types of background knowledge (bk) about the MSA dataset. The bk includes the fact that certain pattern of values in published data is more likely to be observed than other values. For example, this knowledge can be fingerprint correlation (fcorr) knowledge, QI knowledge (qik) [10], or nonmembership knowledge (nmk) [21,22]. MSA values in a table that belongs to a specific individual form a fingerprint. The fcorr between two k-anonymous [4] groups can increase an adversary knowledge. The qik is the personally identifiable information (PII) [21] for an adversary to uniquely identify an individual, and according to nmk, an individual cannot be linked to a specific sensitive value (SV). The (p,k)-Angelization [22] is a strong privacy algorithm for MSAs, where p represents the different sensitivity level of categorical SAs and k implies the k-anonymous QIs. The (p,k)-Angelization algorithm shown in Tables 2 and 3 are obtained from the original microdata in Table 1. The authors in [22] overcame the problem of the nmk attack but still privacy could be breached with fcorr named as the fcorr attack. The privacy breach scenario I explains the invalidation for [22] in detail. The complexity, lack of utility, and privacy breaches in SLOMS [24] and SLAMSA [25] techniques have already been invalidated by the (p,k)-Angelization. Although [22] is an efficient solution for utility improvement, the intruder can easily breach the privacy for a record using the bk and his intelligence. Our work has been motivated by the following limitations in the (p,k)-Angelization algorithm:

(i) Privacy breach scenario I. For example, an adversary (i.e., David) intends to identify p2 (Lisa) information in Table 1. Since they both live in a neighbourhood, age, gender, and Zipcode are known (21, F, and 34607). Using QIs, David identifies her presence in group 3 of the generalized table (GT), i.e., Table 2, and through the batch ID, the sensitive batch table (SBT), i.e., Table 3, in group 3 can be accessed. For the (p,k)-Angelization, physician is a maximum weighted attribute (see Section 5.2). The maximum weighted attribute implies high dependency that has high privacy risk. An attack on it can easily breach privacy. So the intruder starts the attack from the physician attribute. It is an iterative process that leads to the record identification of the target

| Personal attribute | Quasi-identifier attributes QIDs | Cancer type | Cancer type | Diagnosis date | Symptom | Diagnostic method |
|--------------------|---------------------------------|-------------|-------------|----------------|---------|------------------|
| Name               | Gender  Age  Zipcode            | Treatment   | Physician   | Date           |         | Method           |
| p1 (Michael)       | M       34  34548               | Rectal      | Jack        | 7/8/19         | Back pain| Chest X-ray      |
| p2 (Lisa)          | F       21  34607               | Breast      | Alan        | 17/8/19        | Swelling | Ultrasound       |
| p3 (Richard)       | M       26  37506               | Colon       | Daisy       | 22/11/19       | Back pain| Blood test       |
| p4 (Dave)          | M       31  34549               | Prostate    | Tom         | 29/08/19       | Abdominal pain | Chest X-ray |
| p5 (Kate)          | F       35  33753               | Prostate    | Radiation   | 9/09/19        | Testis swelling | Blood test       |
| p6 (William)       | M       38  43674               | Liver       | Ablation    | 5/08/19        | Weight loss | CT scan          |
| p7 (Robert)        | M       27  35064               | Rectal      | Medication  | 12/08/19       | Abdominal pain | CT scan          |
| p8 (Olivia)        | F       32  44662               | Prostate    | Biologic    | 21/09/19       | Back pain | Blood test       |
| p9 (Emily)         | F       22  34548               | Breast      | Biologic    | 13/09/19       | Skin imitation | MRI test         |

Table 1: Original data table T.

| | Gender | Age | Zipcode | Batch ID |
|-----------------|-------|------|---------|----------|
| Person          | 22–26 | 34548−37506 | 1       |
| Person          | 31–38 | 34549–44662 | 2       |
| Person          | 21–34 | 34548−34607 | 3       |
| Person          | 27–35 | 33753–35064 | 4       |

Table 2: Generalized table (GT) from the (p,k)-Angelization.
SAFB 1 and SAFB 3. Although there is no common value for the diagnostic method and the values in SAFB 3 are only related to rays, the intruder is intelligent enough that he can easily assign MRI test to Alan. This may not be the exact value, but can help to guess or identify the record. For symptom although we already have swelling for Alan, taking back pain for Alan which is already assigned to Jack and the only two values in this cell do not suit to the intruder knowledge. For cancer treatment, the common value is surgery and for the cancer type breast is the only attribute value. In SAFB 3, the leftover values are Rectal for Jack and in SAFB 1 and colon for Daisy. At the end, as the intruder also knows that p2 is a female, her attribute values {breast cancer, swelling, and ultrasound/MRI} can easily identify p2. The weighted sensitive attributes values disclosed against the linkable (S) SA identifies the patient p2 record. Table 4 shows that during the process, the details about patient p1 and p4 are also identified. Some of the information regarding Frank and Daisy is incomplete or incorrect due to the fact that no further intersection with any other group is possible since the current data are sample data. This process iteratively executes and can also identify the remaining patients MSAs values.

(ii) Need for bucketization. Deeply analysing the (p, k)-Angelization, it is observed that all the features of angelization [16] were not well utilized. Tables 2 (GT) and 3 (SBT) by the (p, k)-angelization have one-to-one correspondence/linking between the two tables using the bucket id (BID). Due to the one-to-one correspondence, both tables are not considered as independent while the purpose of angelization was to publish both tables independently. Applying SA diversity may affect the utility in GT. Similarly, increasing the dimensionality in QIs in GT also decreases the utility. The adversary after finding a presence of an individual in a bucket in GT can easily move from GT to the exact group in SBT, where the S fingerprint buckets may help in isolating the sensitive values. In fact, splitting the table into GT and SBT in the (p, k)-angelization is useless.

In our proposed (c, k)-anonymization algorithm, the bucketization approach is adopted, which separates the QIs and SAs into two separate tables: generalized table (GT) and sensitive table (ST), independently. Both tables are respectively linked through BID.

GT consists of k-anonymous QIs generalized buckets (GBs), and an adversary cannot get additional information about an individual’s privacy. The ST is a bucket table with
MSAs in the bucketized form named as sensitive attributes fingerprint buckets (SAFBS). Anatomy [26] and Angel [23] are examples of bucketization for preserving privacy; however, they are applicable to single sensitive attributes. In this work, we use the bucketization for MSAs that can prevent different types of adversary’s attack, e.g., fcorr attack. Tables 5 and 6 are the GT and ST produced by the proposed \((c, k)\)-anonymization algorithm. In the proposed approach, better privacy has been achieved with minimum utility loss. It is also not necessary that the publisher should always publish the data with all their QIs attributes known as marginal publication. Marginal publication is to publish the GT with few QI attributes instead of all QI attributes, along with ST. k—_he idea of marginal publication was introduced in [23]. k—_he bucketization has the minimum information loss because of the independent publishing of the GT and ST. In both these tables, the connection is not between the buckets, instead it is between the records in generalized buckets and sensitive buckets.

### 1.2. Contributions

We propose an efficient solution \((c, k)\)-anonymization for privacy preservation in MSAs. In \((p, k)\)-angelization [22], privacy can be breached under the fcorr attack (explained in Section 1.1). The tables published by the proposed \((c, k)\)-anonymization are depicted in Tables 5 and 6. The “Name” attribute in Table 5 is not published while publishing the data. The proposed approach also prevents against the adversary nmk and qik. The main contributions are as follows:

(i) We propose an improvement of \((p, k)\)-angelization, namely as the \((c, k)\)-anonymizaition algorithm, for MSAs privacy. The proposed solution prevents against fcorr attack. For reducing the privacy risk, the real (i.e., one to one) linking between GT and ST is transformed to one-to-many (i.e., real and likely) linking.

(ii) We formally model and investigate the invalidation of \((p, k)\)-angelization for the fcorr attack and correctness of the proposed \((c, k)\)-anonymization algorithm.

(iii) Based on the above points, the experimental results prove that our proposed approach provides better privacy and utility as compared to its counterpart.

### 2. Related Work

In this section, we broadly categorize the data privacy models in order to define boundaries of the proposed work in the available literature.

#### 2.1. Data Privacy Models and Methods

Privacy models can be categorized as (i) syntactic (i.e., partition), or (ii) semantic (i.e., randomized). The syntactic approach achieves privacy in two levels: clustering data and privacy framework. The \(k\)-anonymity
[1] and then its extension l-diversity [3] and then the t-closeness [4] are the examples of syntactic data privacy models, in which the final set of groups are called equivalence classes (ECs). In the semantic approach, the original values are noised in a random way. ε-differential privacy [27] is an example of the semantic data model. The researchers have proposed both the syntactic and semantic privacy models for different types of data, e.g., single sensitive attribute [3, 4, 6], or MSAs [7–9], or 1 : m (i.e., one individual having many records) [28] microdata. For preserving the privacy, the algorithms in privacy models practice different approaches. These approaches can be categorized to (i) generalization [1–5, 13–15] (i.e., greedily convert the more specialized values to less specialized values), (ii) anatomy [25, 26] (i.e., partition the QI and S attributes), and (iii) microaggregation [29, 30] (i.e., dataset is partitioned into clusters where QI values of records are replaced with the mean of value). The proposed work in this paper considers the syntactic data privacy, using generalization and anatomy for MSAs.

2.2. Syntactic Anonymization Literature for Multiple Sensitive Attributes. A plethora of research contributions related to MSAs privacy [7–14, 16–18, 22, 24, 25, 31–38] exists. A recent work, anatomization with slicing [25] is an effective technique for MSAs. Although it does not generalize the QIs attributes, it enhances utility but publishes many tables, which makes the solution more complex. To prevent against proximity breach, the authors in [7] have adopted multi-sensitive bucketization (MSB) technique using clustering. However, it is applicable to numerical data only. The (α, l) model [8] for a single sensitive attribute satisfies the privacy requirements for MSAs. The authors in [31] prevent the negative and positive disclosure of associating between MSAs. In [33], rating of MSAs was proposed that fulfills the privacy requirements. However, the inherent relationship between the SAs can cause association rule attack. An adversary can use related bk to breach the privacy. The authors in [32] prevented the data from association attack and removed the weakness of the rating algorithm. In [37, 38], the authors perform vertical partitioning (i.e., anatomy) and implement decomposition and decomposition plus, respectively, to achieve l-diversity for MSAs. Decomposition plus [38] optimizes the noise value selection in [37] and keeps it closer to the original. The possibility of skewness and similarity attacks in [4, 39] was eliminated by the p’ sensitive t-closeness model [40]. It combines the good features from p-sensitive k-anonymity [39] and t-closeness [4] approaches.

ANGELMS (anatomy with generalization for MSA) [34] vertically partitions the dataset into the QIs table and several SAs tables satisfying the k-anonymity [1] and l-diversity [3] principles, but still it can be attacked with similarity, skewness, and sensitivity attacks. In [16, 18], the KC Slice for dynamic data publishing of MSAs integrates the features of KC-privacy and slicing techniques. The authors have presented the method for a single release, and no studies for multiple releases are available to prove the dynamic claim. In [35, 36], MSAs were handled for achieving privacy but the l-diversity [3] principle was directly adopted that caused huge information loss. The nmk attack was prevented in [41] but still caused high information loss due to the grouping conditions over the data and vulnerability to the background join attack.

The proposed work categorizes the sensitivity of MSAs as top secret, secret, less secret, and nonsecret. c-diverse fingerprint buckets are created that contain records from different categories. The QI values of the created fingerprint buckets are bottom-up generalized through k-anonymity [1].

3. Preliminaries

Let table $T = (\{EI, QI, S\}$ (as shown in Table 1) is the private data form for a publisher to publish. Let there be $t$ tuples in $T$, and each tuple represents an individual or record respondent $i$. The components for tuple $t ∈ T$ are explicit identifier attributes (also called identifying attribute) $EI = \{A_{1}^{\text{EI}}, A_{2}^{\text{EI}}, \ldots, A_{m}^{\text{EI}}\}$, quasi identifier attributes (also called partial identifiers) $QI = \{A_{1}^{\text{QI}}, A_{2}^{\text{QI}}, A_{3}^{\text{QI}}, \ldots, A_{d}^{\text{QI}}\}$, and sensitive information attributes $S = \{A_{1}^{\text{S}}, A_{2}^{\text{S}}, A_{3}^{\text{S}}, \ldots, A_{n}^{\text{S}}\}$. QIs are the partial identifiers or personally identifiable information (PII) that can identify an individual if linked with external data, e.g., voting or census data. Data privacy is all about protecting the sensitive information, which are the confidential and private information belonging to an individual. In this work, we consider a challenging scenario of more than one sensitive attributes for a single individual named as multiple sensitive attributes (MSAs). Notations used in the paper are shown in Table 7.

| Table 6: $(c, k)$-anonymization sensitive table (ST). |
|------------------------------------------------------|
| **Physician** | **Cancer type** | **Cancer treatment** | **Diagnosis method** | **Symptom** | **Diagnosis date** | **Bucket ID** |
|--------------|----------------|----------------------|---------------------|------------|-------------------|--------------|
| Jack, Alan | Rectal, breast, prostrate | Surgery, inhibitor therapy, biologic therapy | Chest X-ray, ultrasound, blood test, MRI test | Back pain, swelling, skin imitation | 7/8/19, 17/8/19, 21/09/19, 13/09/19 | 1 |
| Daisy, Tom | Colon, liver | Surgery, ablation | Blood test, CT scan | Back pain, weight loss | 22/11/19, 5/08/19 | 2 |
| Tom, Frank | Prostate, rectal | Biologic therapy, radiation, medication | Chest X-ray, blood test, CT scan | Abdominal pain, testis swelling | 29/08/19, 9/09/19, 12/08/19 | 3 |

Definition 1 (MSA fingerprint [22]). The MSAs values in table $T$ that belongs to a specific individual form a fingerprint known as MSA fingerprint.

Definition 2 (sensitive attribute fingerprint bucket (SAFB)). A sensitive attribute partitioning of the microdata $T$ consists of a list of SAFB: $\{FB_{1}, FB_{2}, \ldots, FB_{m}\}$ according to the following conditions:
by FB correlation linked through the high-weighted SA value. Each FB contains few fingerprints that belong to k individuals in a specific GB inside GT. The adversary uniquely identifies an individual from the fingerprint correlation knowledge which has direct correspondence with the QI values.

Definition 4 (nonmembership knowledge (nmk) [22]). If an adversary knows that an individual i in GB cannot be linked to a specific SV in FB it is known as nmk.

Definition 5 (fingerprint correlation knowledge (fcorr)). The MSA values obtained from correlating two linkable FBs, i.e., FBj ∩ FBj = Ø can be assigned to a specific individual.

Based on the available information, we consider the adversary’s bk consists of bk = {GT, ST, ED, nmk, qik, fcorr}, where

(i) GT (generalized table) = {A1, A2, A3, ..., A_k, BID}
(ii) ST (sensitive table) = {A1, A2, A3, ..., A_m, BID}
(iii) Any external dataset ED = {ED1, ED2, ..., ED_d} available publically

The adversary applies the bk on available anonymized data to perform an attack and to breach an individual’s privacy.

Definition 6 (fingerprint correlation (fcorr) attack). The adversary with known QIs values and fcorr is able to perform fcorr attack by deducing single SVs from the intersection of FBs that are linked via δ. The fcorr attack can be

(i) Partial (pfcorr) attack: few of the SAs from fingerprints in two or more FBs may produce unique SVs
(ii) Full (ffcorr) attack: all of the SAs from fingerprints in two or more FBs may produce unique SVs

For an adversary, the ffcorr attack has no doubt to uniquely identify an individual while with the pfcorr attack can identify a record respondent if the resulted sensitive information belongs to the above minimum weighted attributes (ζ_w) (see 5.1 algorithm). The ζ_w attributes do not contribute in an individual record identification.

Definition 7 (high-level petri-nets (HLPNs) [44]). A petri-net is used as a model to examine the control of information in a system. HLPN formally analyses the system with mathematical properties. A HLPN is a 7-tuple N = {P, T, F, φ, R_n, L, and M_0}, where P is a set of places represented by circles, T is a set of transitions represented by rectangular boxes such that P ∪ T ≠ Ø and P ∩ T = Ø, F is the flow relations such that F ⊆ (P × T) ∪ (P × T), L are the labels on F, φ maps places to types., R_n represents the rules for transitions, and M_0 is the initial marking. In short, L, φ, and R_n represent the static semantic, whereas P, T, and F depict the dynamic structure.
4. Critical Review for \((p, k)\)-Angelization with \(fc\text{corr}\) Attack Identification Using Formal Modelling and Analysis

**Definition 8** \((p, k)\)-Angelization [22]). A pair of bucket partitioning \(=\{B_1, B_2, B_3, \ldots, B_m\}\) and batch partitioning \(=\{C_1, C_2, C_3, \ldots, C_p\}\) of table \(T\) produces two tables \(GT\) and \(SBT\) such that

(i) GT consists of QI attributes with batch id (BID) belonging to table \(T\). The QI values from \(t\) records are \(k\)-anonymously grouped and linked with SBT via BID.

(ii) SBT consists of (SA, BID), where BID is \(i (1 \leq i \leq m)\) and SA are the MSA from table \(T\).

(iii) Batch partitioning satisfies \((p, k)\)-anonymity [16] where every bucket has records from \(p\) categories and have \(k\)-tuples to prevent against linking attack while the group partitioning satisfies \((p, k)\)-anonymity [23].

The following reasons explains the invalidation for \((p, k)\)-angelization [22].

(i) The invalidation of the existing \((p, k)\)-angelization is due to the \(fc\text{corr}\) that causes \(fc\text{corr}\) attack (as shown in Table 4). Although the records from \(p\) categories are \(k\) indistinguishable on MSAs fingerprint, they are uniquely distinguishable because of unique SAFB values obtained from linkable buckets. So, Lemma 2 in [22] is incorrect. Its corrected form is given in Lemma 1 (Section 5).

(ii) Invalidation of theorem 2 in [15]: the adversary can correlate the sensitive information from qik. Scenario I explains the \(fc\text{corr}\) attack that extracts unique sensitive information using QI values.

Now, we formally model the \((p, k)\)-angelization algorithm to check its invalidation with respect to the \(fc\text{corr}\) attack. The \((p, k)\)-angelization algorithm is depicted with HLPN and formally analysed with its mathematical properties. The purpose of using HLPNs is to depict (i) the interconnection of the model components and processes, (ii) a clear flow of data among the processes, and (iii) the in depth inside about how the process of information takes place, in order to isolate the flaw in \((p, k)\)-angelization. Figure 1 depicts HLPNs for \((p, k)\)-angelization. The variable types and mapping of data types on places are shown in Tables 9 and 10, respectively. The adversarial model in Figure 1 comprises of three entities: end user, trusted data sanitizer, and adversary. The initial transition is referred to as input transition that contains the raw data (e.g., patient’s EHRs) collected from a health organization. The trusted data sanitizer anonymizes the data using the \((p, k)\)-angelization algorithm and produces GT (Table 2) and SBT (Table 3) tables. The produced tables are ready to be published which are exploited by the adversary through the \(fc\text{corr}\) attack in Table 4.

Rule 1 checks the existence of the number of dependent SAs with respect to another SA. Rule 2 counts the dependent attributes and selects the maximum weighted attributes. However, if there exists more than one in the weight set, then an external factor \(\varepsilon\) is added to one of them, based on some external facts. The weight set is sorted in descending to select the maximum weight as in rule 3 and 4. Based on weight calculation and MSAs in \(T\), the category table is formed in rule 5.

| Category ID | Physician | Cancer type | Cancer treatment | Diagnosis method | Symptom | Diagnosis date | Sensitivity level |
|-------------|-----------|-------------|------------------|------------------|--------|----------------|------------------|
| One         | Daisy     | Colon, prostate | Surgery, radiation | Blood test | Back pain, testis swelling | 22/11/19, 9/09/19 | Top secret |
| Two         | Jack      | Rectal, prostate | Surgery, biologic therapy | Chest X-ray, blood test | Back pain | 7/8/19, 21/09/19 | Secret |
| Three       | Alan      | Breast       | Inhibitor therapy, biologic therapy | Ultrasound, MRI test | Swelling, skin irritation | 17/8/19, 5/08/19, 12/08/19 | Less secret |
| Four        | Tom       | Prostate, liver, rectal | Biologic therapy, ablation, surgery | Chest X-ray, CT-scan | Weight loss, abdominal pain | 29/08/19, 5/08/19, 12/08/19 | Nonsecret |

Table 8: Category table from original microdata table \(T\).
Table 9: Types used in HLPN for the \((p, k)\)-anangelization.

| Types   | Descriptions                                                                 |
|---------|-------------------------------------------------------------------------------|
| \(T\)   | Place holding integer and string type data                                    |
| PID     | An integer identifying patient id                                             |
| \(A^{DS}\) | A set consists of dependent sensitive attributes                      |
| \(w\)   | An integer defines final calculated weights of \(A^{DS}\)                    |
| \(w_{\text{sort}}\) | Set containing sorted weight of SA in descending                          |
| \(w_{\text{max}}\) | Maximum weight after adding external factor \(\chi\)               |
| C       | Category Id for MSA                                                           |
| BID     | Batch id used to connect GT and SBT                                           |
| QI      | Quasi-identifier for ith end user                                            |
| MSF     | Multiple sensitive fingerprint values for ith end user                        |

Table 10: Mapping of data types on places.

| Places | Mapping                                                                 |
|--------|------------------------------------------------------------------------|
| \(\phi(T)\) | \(P \times (\text{QI} \times \text{MSF} \times \text{PID})\)          |
| \(\phi(A^{DS})\) | \(P \times (A^{DS})\)                                           |
| \(\phi(w)\)   | \(P \times (\text{w})\)                                           |
| \(\phi(w_{\text{sort}})\) | \(P \times (w_{\text{sort}})\)                                 |
| \(\phi(w_{\text{max}})\) | \(P \times (w_{\text{max}})\)                                 |
| \(\phi(\text{CtgT})\) | \(P \times (\text{C} \times \text{MSF})\)          |
| \(\phi(\text{GT})\) | \(P \times (\text{QI} \times \text{BID})\)                 |
| \(\phi(\text{SBT})\) | \(P \times (\text{MSF} \times \text{BID})\)                 |
| \(\phi(\text{qik})\) | \(P \times (\text{QI} \times \text{BID})\)                 |
| \(\phi(\text{fcorr})\) | \(P \times (\text{MSF} \times \text{BID})\)                 |
| \(\phi(\text{MSA Dis})\) | \(P \times (\text{QI} \times \text{MSF} \times \text{PID})\) |

Rule 6: \(R(\text{pk – Angel}) = \forall i12 \in x12, i13 \in x13, i14 \in x14, i15 \in x15\)
\[i14[1] := \text{angel}((i12[1]), [i13[1], i13[2], i13[3]])\]
\[\land i14[2] := \text{bucketID}(i12[1])\]
\[\land i14' := x14 \cup ([i14[1], i14[2]])\]
\[i15[1] := \text{angel}((i12[1]), [i13[1], i13[2], i13[3]]) \land i15[2] := \text{bucketID}(i12[1])\]
\[\land i15' := x15 \cup ([i15[1], i15[2]])\]

Rule 7: \(R(\text{fcorr Attack}) = \forall i16 \in x16, i17 \in x17, i19 \in x19, i21 \in x21, i22 \in x22\)
\[\land qik\in [i16[1], i19[1]] \longrightarrow i22[1] \land i13[1] \land \text{fcorrDis}(i17[1], i17[1], i21[1]) \longrightarrow i22[2] := i13[2] \land i22[3] := i13[3]\]

5. Proposed \((c, k)\)-Anonymization for Multiple Sensitive Attributes

Although the \((p, k)\)-Angelization model is a state-of-the-art approach for MSAs, especially for the categorization of sensitive values. But the ST still lacks in privacy because of blindly using the same angelization approach for MSAs. It leads to fcorr attack, i.e., fcorr attack or pfcorr attack. We name the improved form of \((p, k)\)-anangelization, \((c, k)\)-anonymization for MSAs and is describe as follows:

**Definition 9 ((c, k)-anonymization).** A pair of generalized bucket partitioning \(\{B_1^c, B_2^c, \ldots, B_n^c\}\) and sensitive attribute fingerprint bucket partitioning \(\{FB_1, FB_2, \ldots, FB_m\}\) of table \(T\) produces two tables \(GT\) and \(ST\) such that:

1. \(GT\) consists of QI attributes buckets \(B_i^c\) \((1 \leq i \leq n)\) with bucket id (BID) belonging to table \(T\). The QI values belonging to \(t\) records are \(k\)-anonymously
grouped and linked with corresponding sensitive buckets in ST via BID. And $\bigcup_{i=1}^{n} B_i^T = \text{GT}$, and for any $i \neq j$, $B_i^T \cap B_j^T = \emptyset$.

(ii) ST consists of (FB, and BID), where BID is $i$ $(1 \leq i \leq n)$ and FB are the MSAs in buckets from Table T. And $\bigcup_{i=1}^{n} B_i^T = \text{ST}$, for any $i \neq j$, and $B_i^T \cap B_j^T = \emptyset$ among linkable $QI$ buckets through maximum weighted sensitive attributes ($\delta$).

(iii) Generalized bucket partitioning satisfies definition 3 and sensitive bucket partitioning satisfies definition 2 based on the category table $\text{CtgT}$ (Table 8). Every generalized bucket has $k$-tuples to prevent against linking attack. The sensitive partitioning has $c$-diverse records from $c$ categories in $\text{CtgT}$ such that

(a) for strict approach, fulfill equation (4)
(b) for relax approach, fulfill equation (5)

Lemma 1 (uncertainty in $\mathcal{L}$ SAFBs). If for T having the MSAs dataset, the anonymized form $T'$ satisfies ($c$, $k$)anonymization, then $T'$ satisfies ($c$, $k$)-diversity for MSA fingerprints.

Proof. Let sbp be the random sensitive bucket partitioning in $T'$. There must be at least $t$ tuples from $c$ categories in sbp such that $(\text{sbp}_a \cap \text{sbp}_b > 0 \text{ or } 1, a \neq b)$ or the $fcorr \leq \zeta_{w_j} \bigvee \text{sbp}_x$, where $\zeta_{w_j}$ are the minimum weighted SAs having the lowest dependency (see algorithm 5.1). So the $c$ categorized records are indistinguishable on the MSA finger-

5.1. Privacy Risk: Feasibility of the Proposed Work. The presence attack and fcorr attack are the two possible attacks that breach the privacy of GT and ST, respectively. The adversary can breach an individual’s privacy by linking the obtained sensitive information with the QI values and with the bk. Every QI record has a corresponding SA fingerprint in a specific FB, as depicted in Tables 2 and 3. The $(p, k)$-angelization has the one-to-one real linking through BID. The real linking has high privacy leakage and 100% chances of presence attack. Another type of linking can be the likely linking between GT and ST where all BID linking are not real. Relating the real and likely linking, the privacy risk for an individual is defined as PriRisk$(t) = (r/m)$, where $r$ are the real linking and $n$ are the total number of likely linking. The likely linking for a specific size of EC varies and it depends on the QI values in GT that have linking with certain FBs. For preventing privacy leakage PriRisk$(t) \leq (1/l)$, where $l$ represents the $l$-diversity [3]. Every FB has $c$-diverse finger-

Consider the privacy risk for the given Tables 5 and 6 processed by proposed ($c$, $k$)-anonymization. In GT Table 5, for $c=2$, there are three different sizes of ECs that have varying $l$-diversity (ranging $2 \leq l \leq 4$) in ST Table 2. For EC size 4, $r = 4$ and $n = 4 \times 4 + 4 \times 2 + 4 \times 3 = 42$; so, PriRisk$(t) = (r/n) = (4/42) \leq (1/4) \Rightarrow 0.09 \leq 0.25$, which is a very low privacy risk. Similarly, for EC size 2, $r = 2$ and $n = 2 \times 4 + 2 \times 2 = 12$; so, PriRisk$(t) = (2/12) \leq (1/2) \Rightarrow 0.16 \leq 0.5$. For the last EC size 3, $r = 3$ and $n = 3 \times 3 + 3 \times 4 = 21$; so, PriRisk$(t) = (3/21) \leq (1/3) \Rightarrow 0.14 \leq 0.3$. Even in case when in an EC, the minimum distance QI values link to a single FB in ST, for example, when EC size is 2, then $r = 2$ and $n = 2 \times 2 = 4$ so PriRisk$(t) = (2/4) \leq (1/2) \Rightarrow 0.5 \leq 0.5$, and the probability of privacy disclosure is the diversity in the FB. In certain cases, the higher utility in QI values may further increase the likely linking which will more reduce the privacy risk. This proves that the proposed approach has very low privacy risk for the presence attack and data disclosure.

5.2. ($c$, $k$)-Anonymization Algorithm. The objective of the proposed ($c$, $k$)-anonymization algorithm is to provide a sustainable privacy for MSAs. The algorithm gets a microdata table T (Table 1) as input and produces two anonymized tables, i.e., GT (Table 5) and ST (Table 6).

The proposed Algorithm 1 performs two major functionalities: categorizing the MSAs based on calculated weights and creating secure FBs for the whole dataset. For SA categorization, the algorithm calculates weights for all the MSAs in the dataset to get to know the dependency of SAs. The dependency shows the sensitivity level of the SAs. These weights are sorted in the descending order to get categorized MSAs. The weight calculation for MSAs creates the category table (CtgT) that helps to create $l$-diverse ($c$-diverse with respect to category) FBs. The FBs must satisfy equations (4) and (5) in order to prevent fcorr attack. Therefore, if some of the FBs are not according to equations (4) or (5) they are refined to fulfil. The purpose to refine is because the input data may be of different nature and may contain SVs that may not be grouped initially. The complete algorithm (Algorithm 1) and its working are explained in the following:

Sensitive attributes weight calculation. In Algorithm 1, a calling function $\text{wtCalc}(\cdot)$, shown in Function 1, calculates weights for all the MSAs from Table 1. There are six different types of SAs, and each has its own level of sensitivity or sensitivity weights. Let $W = \{w_1, w_2, \ldots, w_n\}$ is the set of weights for each SA such that $s_1$ has weight $w_1$, $s_2$ has weight $w_2$, and so on. SA weights are the dependency on all other SAs. Similar to [22] the weight is calculated in the following equation:

$$w_i = \sum_{j=1}^{m} \text{Dep}(s_j, s_i),_{\forall \text{SA } \text{and } v}$$

where $m$ are the total number of attributes dependent on attribute $s_i$. The dependency of an SA is determined by the total range of attributes identifying the SAs. The for loop in the beginning calculates the sensitivity for $s_i$ with all other SAs, i.e., $s_j$. This determines the sensitivity level for all the SAs. To calculate the weights (second for loop), cardinality checking is performed of all dependent attributes. Calculated weights for the SAs that exist in microdata are shown in Table 11.
Maximum weighted sensitive attributes selection. From the calculated dependencies through Function 1, maximum weighted attributes $\xi_w$, are selected. Maximum weights mean high dependency leads to high disclosure risk. So attributes set with $\xi_w$ needs maximum protection. Although there are very rare chances, the problem arises if there exists more than one SAs having the same $\xi_w$. Then, an external factor $\chi$ is added to select only one maximum weighted attribute. The algorithm adds an external factor $\chi$, i.e., $w_i + \chi(w_j)$ to each attribute in set $\xi_w$, and the weights $w_i$ are then sorted in the descending order to get one single maximum weighted SA ($\delta$) (can be seen in Algorithm 1).

Categorizing sensitive attributes. Attribute occurring in the first location of the set $w_i$ is selected as $\delta$. The descending order of calculated weights for MSA is categorized through the categorize() function as top secret, secret, less secret, and nonsecret, in Table 8. From weight calculations, although disease and physician have equal weights, we select physician as the maximum weighted SA ($\delta$) because of some other external factors $\chi$. For example, physician information may provide a reason for an attack. Therefore, to minimize the $L'$, we use the linkability control factor $\epsilon_f$, $1 \leq \epsilon_f \leq \min\text{count}$ for specific $\delta$). $\epsilon_f$ minimizes the repetition of the same $\delta$ value in different FBs. The table (Table 6) published by the proposed algorithm has $\epsilon_f = 2$, which means that a maximum of two records for a single physician can exist in an FB. This brings the existing $\delta$ values to minimum FBs and $L'$ is reduced. So, the chances of the fcorr attack on possible FB are ultimately reduced (Table 3 has 3 $L'$ FBs while Table 6 has only 1 $L'$ FB). The high value for $\epsilon_f$ further reduces $L'$

### Algorithm 1: $(c, k) –$ anonymization.

**Input:**
- $T$: Microdata table = [ID, QI, S]
- $\chi$: External Factor

**Output:**
- GData: Generalized Table-GT
- SAFB: Sensitive Table-SY

Procedure CK - ANONYMIZATION(T)

**Input:**
- SAFB: [ ], GData = [ ]
- $w_i = wtCalc(w, D, s, h)$
- $\forall i \in \xi_w = (\max(w_i), \min(w_i))_{i \in w}$
- $\forall w_i \in \xi_w$, $w_i = w_j + \chi(w_j)$
- $v = Sort(w_i)$
- $\delta = (\max(v))$

Ctg = Categorize($\delta$, MSA)

FB = CreateFB(CtgT, T)

while FB $\neq $ {
  if FB$_k$ satisfy equation (4) then
    SAFB = SAFB $\cup$ FB$_k$
  else
    SAFB $\cup$ RefineFB()
  end if
end while
GData = Generalization(T, SAFB, k)
return GData, SAFB

**Function 1: Function wtCalc($w_i$, Dep, $s$, and $h$).**

| Sensitive attributes | Identified by | Dependency | Weightage |
|-----------------------|--------------|------------|-----------|
| s1: cancer type       | s2, s3, s6   | 3          | 3         |
| s2: cancer treatment  | s1           | 1          | 1         |
| s3: physician         | s1, s2, s6   | 3          | 3         |
| s4: diagnostic date   | —            | 0          | 0         |
| s5: symptoms          | —            | 0          | 0         |
| s6: diagnostic method | s1           | 1          | 1         |

Maximum weighted sensitive attributes selection. From the calculated dependencies through Function 1, maximum weighted attributes $\xi_w$, are selected. Maximum weights mean high dependency leads to high disclosure risk. So attributes set with $\xi_w$ needs maximum protection. Although there are very rare chances, the problem arises if there exists more than one SAs having the same $\xi_w$. Then, an external factor $\chi$ is added to select only one maximum weighted attribute. The algorithm adds an external factor $\chi$, i.e., $w_i + \chi(w_j)$ to each attribute in set $\xi_w$, and the weights $w_i$ are then sorted in the descending order to get one single maximum weighted SA ($\delta$) (can be seen in Algorithm 1).

Categorizing sensitive attributes. Attribute occurring in the first location of the set $w_i$ is selected as $\delta$. The descending order of calculated weights for MSA is categorized through the categorize() function as top secret, secret, less secret, and nonsecret, in Table 8. From weight calculations, although disease and physician have equal weights, we select physician as the maximum weighted SA ($\delta$) because of some other external factors $\chi$. For example, physician information may provide a reason for an attack. Therefore, to minimize the $L'$, we use the linkability control factor $\epsilon_f$, $1 \leq \epsilon_f \leq \min\text{count}$ for specific $\delta$). $\epsilon_f$ minimizes the repetition of the same $\delta$ value in different FBs. The table (Table 6) published by the proposed algorithm has $\epsilon_f = 2$, which means that a maximum of two records for a single physician can exist in an FB. This brings the existing $\delta$ values to minimum FBs and $L'$ is reduced. So, the chances of the fcorr attack on possible FB are ultimately reduced (Table 3 has 3 $L'$ FBs while Table 6 has only 1 $L'$ FB). The high value for $\epsilon_f$ further reduces $L'$

| Sensitive attributes | Identified by | Dependency | Weightage |
|-----------------------|--------------|------------|-----------|
| s1: cancer type       | s2, s3, s6   | 3          | 3         |
| s2: cancer treatment  | s1           | 1          | 1         |
| s3: physician         | s1, s2, s6   | 3          | 3         |
| s4: diagnostic date   | —            | 0          | 0         |
| s5: symptoms          | —            | 0          | 0         |
| s6: diagnostic method | s1           | 1          | 1         |
but increases information loss, so a balance should be maintained between $c_f$ and utility preservation.

Refining FBs. FBs are refined through the function RefineFB(), as depicted in Function 3. The fcorr attack between any two $\mathcal{L}$ FBs can breach the privacy. The purpose is to completely avoid the correlation. A percentage of record that discloses from intersection between $\mathcal{L}$ FBs can be associated to a specific individual. For example, any percentage of record obtained from equation (3) results in a single value for each SA correlation is a privacy breach and is not acceptable, especially for high weighted attributes.

$$\text{FB}_i \cap \text{FB}_j = \{A_{ji} \cap A_{ji}^1, A_{ji}^1 \cap A_{ji}^2, \ldots, A_{ji}^a \cap A_{ji}^b\},$$ \hspace{1cm} (3)

where $i \neq j$ and $i, j \in \mathcal{L}$ via $\delta$. The high percentage disclosure infers a high intruder confidence for privacy breach and vice versa. The decreasing order of $w_i$ is the decrease in probability of privacy breach, among the $n$ $\mathcal{L}$ FBs. The measure to prevent against the fcorr attack is no or minimum data exposure from intersection. The refining process in Function 3 for FBs works under two approaches, i.e., strict and relax. Strict approach is given in the following equation:

$$\text{FB}^{w_i - \delta}_{i,j} \cap \left\{ \text{FB}^{w_i - \delta}_{j} \right\}^n = 0 \text{ or } 1, \ \forall A^z,$$ \hspace{1cm} (4)

\begin{function}
\begin{verbatim}
Input:
FB = \{FB_1, FB_2, FB_3, \ldots, FB_n\}, a set of FBs that are linked via the same $\delta$
$k$: minimum $k$-anonymity level
Output: set of refined FB = \{FB_1, FB_2, \ldots, FB_n\}, linked via same $\delta$.
while FB \# \{\} 
  if N < 2k then
    FB_k = N
    FB = FB \cup FB_k
  else
    r_{xy} = (\forall_{ru, rv, N, c_f, FB_k, ctg_x, ctg_y}) \leq c_f
    // records having same physician
    r_{xy} = \text{distinct} (\forall_{ru, rv, N, c_f, FB_k, ctg_x, ctg_y}) \leq c_f
    // where $u \neq v$ and $x \neq y$
    FB_{xy} = \{r_{xy}\}
    FB = FB \cup FB_{xy}
    N = N \setminus FB_{xy}
  end if
end while
return FB
end function
\end{verbatim}
\end{function}
Input:
- $T$: QI attributes from original microdata $T$
- $k$: $k$ anonymity level
- BID: Bucket identifier from SAFBs
Output:
- GData: $GData = \{\}$
- N = $N_i$ // BID is received from SAFBs
  - sort N w.r.t QIs
  - while $N \neq \emptyset$
    - if $N < 2k$ then
      - GData = GData $\cup$ gen(N) // using BID
    - else
      - GData = GData $\cup$ gen(N)
  end if
end while
return GData
end function

Function 4: Function Generalization($T$, SAFB, $k$).

5.3. Formal Modelling and Analysis for Proposed ($c$, $k$)-Anonymization Algorithm. In this section, we do formal modelling and analysis of the proposed ($c$, $k$)-anonymization algorithm to analyse and formally validate against adversary’s background knowledge, i.e., fcorr attack. We have used HLPN (Definition 7) to model the proposed system. The HLPN provides the mathematical representation to analyse the behaviour of the proposed system. For a system representation in HLPN, first, the data types associated with the $P$ (Places) are defined and then the set of rules for HLPN are defined. Figure 2 represents the HLPN for the proposed ($c$, $k$)-anonymization algorithm. Tables 12 and 13 show the data types and mapping of data types on places that are described which are involved in the proposed algorithm HLPN.

The algorithm begins from weight calculation as in [22] to create the CtgT table (Table 8). So, the transitions in rules 1, 2, 3, 4, and 5 are the same for the proposed ($c$, $k$)-anonymization algorithm. The main goal is to nullify rule 7 and to create such FBs that prevent against fcorr attack. In rule 8, the FBs are created. The function createFB() takes $T$ (i.e., $\Phi$ (QI $\times$ MSF $\times$ PID)), and based on category id $C$ the categorized MSA are accommodated into different $c$ diverse FBs.

Rule 8: $\text{R(Crt FB)} = \forall i12 \in x_{12}, i13 \in x_{13}, i14 \in x_{14} |
i_{14}[1] = \text{createFB}(\{i_{13}[1]\}, \{i_{12}[1], i_{12}[2], i_{12}[3]\})$

$\land i_{14}[2] = \text{bucketID}(i_{13}[2])$

$\land i_{14}' = x_{14} \cup \{i_{14}[1], i_{14}[2]\}$

In the next step, the created FBs in 8 are evaluated to prevent fcorr attack. Rule 9 gets an input of equations (4) and (5) to verify the nonexistence of fcorr knowledge between different $\text{FB}$s. All those FBs that satisfy either of the equation are stored at place SAFB via rule 10, while other SAFBs are forwarded for refinement to satisfy the equations as shown in rule 11.

Rule 9: $\text{R(Chk - Eqs)} = \forall i_{15} \in x_{15}, i_{17} \in x_{17}, i_{18} \in x_{18} |
i_{17}[1]i_{17}[1], i_{18}[1] \rightarrow i_{18}[1] = \text{TRUE} \land i_{18}' = x_{18} \cup \{i_{18}[1]\}$

$\land i_{18}[1] \neq \text{TRUE} \land i_{18}' = x_{18} \cup \{i_{18}[1]\}$

Rule 10: $\text{R(Pick FB)} = \forall i_{19} \in x_{19}, i_{20} \in x_{20}, i_{21} \in x_{21} |
i_{20} = \text{TRUE} \land i_{21}[1] = \text{storeAll}(i_{19}[1], k_{i_{19}}[1] \land i_{21}[2]) \land i_{19} = i_{21} \land i_{21}' = x_{21} \cup \{i_{21}[1], i_{21}[2]\}$

Rule 11: $\text{R(Call RFB)} = \forall i_{22} \in x_{22}, i_{23} \in x_{23}, i_{24} \in x_{24} |
i_{22} = \text{FALSE} \land i_{24}[1] = \text{storeAll}(i_{23}[1], j_{i_{23}[1]}[i_{23}[j]] \land i_{24}[2] = i_{23}[2] \land i_{24}' = x_{24} \cup \{i_{24}[1], i_{24}[2]\}$

The minimum requirement to prevent from the fcorr attack is to at least satisfy equation (5). In rule 12, such requirements are fulfilled via function RefineFB() for all the FALSE FBs from rule 9 and are stored at place SAFB. Rule 13 just combines all the secured FBs from places SAFB and SAFBr to create one ST that can prevent from any bk attack, e.g., fcorr attack.

Rule 12: $\text{R(Crt RFB)} = \forall i_{25} \in x_{25}, i_{26} \in x_{26} |
i_{26}[1] = \text{refineFB}(i_{21}[1], k_{i_{25}[1]}[i_{25}[1]] \land i_{26}[2] = i_{25}[2] \land i_{26}' = x_{26} \cup \{i_{26}[1], i_{26}[2]\}$

Rule 13: $\text{R(Crt ST)} = \forall i_{27} \in x_{27}, i_{28} \in x_{28}, i_{29} \in x_{29} |
i_{29}[1] = \text{combine}(i_{27}[1], i_{28}[1]) \land i_{29}[2] = \text{combine}(i_{27}[2], i_{28}[2]) \land i_{29}' = x_{29} \cup \{i_{29}[1], i_{29}[2]\}$

The generalization of the QIs in T with respect to the sensitivity categorization and the BID obtained from ST (rule 13) is performed in rule 14. The created GT and ST tables via rules 13 and 14, respectively, can thwart against an adversary’s presence and fcorr attacks and are ready to publish. Rule 15 shows the adversary’s zero gain after attack.

Rule 14: $\text{R(Gen)} = \forall i_{30} \in x_{30}, i_{31} \in x_{31}, i_{32} \in x_{32}, i_{33} \in x_{33} |
i_{33}[1] = \text{generalize}(i_{31}[1], i_{30}[1]) \land i_{33}[2] = i_{32}[2] \land i_{33}' = x_{33} \cup \{i_{33}[1], i_{33}[2]\}$

In rule 15, the adversary’s bk, i.e., fcorr, consists of QI, MSA, and PID. To apply the fcorr attack, the adversary compares the SVs in bk with published tables MSA buckets and with the corresponding QIs to get a matching MSA that belong to a specific PID as in the original table T. However, the adversary fails to do so and the union of the bk yields an empty set, which shows that the bk could not identify a specific individual record.

Rule 15: $\text{R(fcorr Attack)} = \forall i_{34} \in x_{34}, i_{35} \in x_{35}, i_{37} \in x_{37}, i_{38} \in x_{38} |
f_{\text{corrDis}}(i_{34}[1], i_{35}[1]) \rightarrow (i_{34}[1], i_{35}[1]) \cup i_{37}[2] \neq i_{12}[2] \land i_{12}[3] \cup i_{38}[1] \cup i_{38}[2] = \emptyset$
code executed on the same machine. The dataset obtained from the Cleveland Clinic Foundation Heart disease is available at https://archive.ics.uci.edu/ml/datasets/Heart+Disease. This dataset consists of 75 attributes. The experiments are performed on two QI attributes: age and gender and 12 sensitive attributes. These attributes are enough to evaluate the performance of the proposed algorithm. Table 14 shows the QIs and SAs used in the experiment with attribute description and number of distinct values in each domain.

Different general purpose posteriori measures for utility and privacy loss [9, 15, 18, 22] are available for generalization-based algorithms. In these approaches, the publisher does not know about the recipient analysing method. The
6.1. Utility Loss. For utility loss measure, we analyse our algorithm using the following techniques.

6.1.1. Normalized Certainty Penalty. NCP measures accuracy loss in the anonymized release. Let $T = \{Q_1, Q_2, \ldots, Q_q\}$ are QI. The information loss for a single QI attribute is $NCP_{Q_i}(t) = \frac{(z_i - y_i)/|Q_i|}{z_i}$, where $y_i = x_i \in T$ is the actual QI value from $T$, and $|Q_i|$ is the domain range on QI, i.e., $\max \{|t.Q_i|, |t.Q_j|\}$ is the number of tuples where $Q_i$ is the domain range on QI. The total weighted certainty penalty for the whole table is the sum of all attributes in a tuple, and then NCP obtained from all tuples is added as shown in the following equation:

$$NCP(T^*) = \sum_{i=1}^{q} \sum_{t \in T^*} w_i \cdot NCP_{Q_i}(t),$$

where $NCP(T) = \sum_{i=1}^{q} \sum_{t \in T^*} w_i \cdot NCP_{Q_i}(t)$ represents penalty for a tuple, $w_i$ are weights associated to attributes, and $T^*$ is the final anonymized release. Figure 3 shows the percentage value for NCP for varying values of $k$-anonymity, keeping fixed number of attributes (e.g., MSA = 6) for analysing (c, k)-anonymization and (p, k)-angelization algorithms. The penalty, i.e., NCP% value for (p, k)-angelization continuously increases while increasing the $k$-anonymity because the $k$ groups have the one-to-one correspondence with FBs in ST. This means high diversity in FBs may further effect the utility in GT. So, the splitting of table $T$ (Table 1) into GT (Table 2) and ST (Table 3) has no benefit at all. The attributes in each table are still dependent on each other. While the proposed (c, k)-anonymization has one-to-many correspondence between ST and GT where the GB creates closer $k$-anonymous groups for the same $k$ size class. The comparatively less generalized QIs have lower utility loss and have almost zero loss.

6.1.2. Query Accuracy or Precision of Data Analysis Queries. The purpose of anonymized data is to extract useful statistics and contribute in decision-making. Such utility of the anonymized release is measured through aggregate query answering. Consider the following type of aggregate query:

$$\text{Select COUNT}(*) \text{ from } T^* \text{ where } A^{Q^1}_1 \in \text{Domain } (A^{Q^1}_1) \text{ AND } \ldots \text{ AND } A^{Q^j}_q \in \text{Domain } (A^{Q^j}_q).$$

Published table $T^*$ has $q$ as $A^{Q^1}_1$, i.e., $A^{Q^1}_1, A^{Q^2}_2, \ldots, A^{Q^q}_q$. The domain $(A^{Q^j}_q)$ size depends on query selectivity ($\theta$) which indicates the expected number of record selection. The selectivity $\theta = |r_Q|/|T|$, where $|T|$ is the total number of records in the dataset and $|r_Q|$ indicates the number of records obtained from query $Q$. To measure the precision

| S. no | Attributes | Description | No. of distinct values |
|-------|------------|-------------|------------------------|
| 1     | QI         | Age         | Age in years (range values) | 48 |
| 2     | Gender     | Male or female | 2 |
| 3     | Cp         | Chest pain type (typ, atyp, non, asymp) | 4 |
| 4     | Trestbps   | Resting blood pressure | 49 |
| 5     | Chol       | Serum cholesterol | 148 |
| 6     | Smoke      | Yes or no | 2 |
| 7     | Cigs       | Cigarettes per day | 24 |
| 8     | Fbs        | Fasting blood sugar (true or false) | 2 |
| 9     | Famhist    | Family history of coronary artery disease | 2 |
| 10    | Restecg    | Resting electrocardiographic results | 3 |
| 11    | Thalach    | Maximum heart rate achieved | 90 |
| 12    | Thalrest   | Resting heart rate | 61 |
| 13    | Cmo        | Month of cardiac cath | 12 |
| 14    | Cyr        | Year of cardiac cath | 4 |
loss, the query error in equation (8) analyses the error between the COUNT queries executed on the published and original datasets.

$$\text{Query error} = \frac{|\text{count (generalized)} - \text{count (original)}|}{\text{count (original)}}.$$  \hfill (8)

Calculating the query error is a common matrix to measure the utility of the anonymized release. We compare the utility of the \((p, k)\)-angelization and \((c, k)\)-anonymization by generating 1000 random queries and averaging their query errors. Figure 4 depicts the query error for \((p, k)\)-angelization and \((c, k)\)-anonymization comparatively. In Figure 4(a), the comparative increase in the query error for varying \(k\) size is because of the new record insertions that increase the generalization range. While the proposed \((c, k)\)-anonymization shows a comparative low error rate because of comparative decrease in QI generalization. The selectivity, i.e., \(\theta\) graph in Figure 4(b), depicts that for high selectivity, more number of records are selected. So, the difference to calculate the query error via equation (8) will automatically decrease.

6.2. Privacy Loss. Identification of individual respondents from an anonymized release is directly proportional to the privacy loss. Therefore, the privacy loss is measured with respect to the number of single record identifications that are obtained from the intersection of \(\mathcal{D}\) FBs, considered as vulnerable records. We analyse the privacy loss with varying value of \(k\) and MSA. Figure 5 shows the results obtained from the experiments for the \((p, k)\)-angelization and \((c, k)\)-anonymization algorithms. In Figure 5(a), for \((p, k)\)-angelization, the number of vulnerable records increases with increase in \(k\) size because there are more number of records that have single SV obtained from the intersection among \(\mathcal{D}\) FBs. Similarly, in Figure 5(b), with increasing the MSAs, the chances of getting a single SV against each SA increases and hence the number of vulnerable records increases for \((p, k)\)-angelization. While in both cases, i.e., Figures 5(a) and 5(b), the proposed \((c, k)\)-anonymization has no such single SV from the intersection in \(\mathcal{D}\) FBs because of satisfying equations (4) and (5). Therefore, there exists no vulnerable records in the proposed \((c, k)\)-anonymization algorithm.

6.3. Execution Time Analysis. The execution time for the proposed \((c, k)\)-anonymization is high as compared to the \((p, k)\)-angelization. This is because of the privacy requirements to satisfy equations (4) and (5). Although the categorization and record selection are almost the same in both algorithms, however, to satisfy equations (4) and (5) is time consuming and may need further time to merge records between \(\mathcal{D}\) FBs. Therefore, at the cost of improved privacy, the execution time increases. In Figure 6, with the increase in number of MSAs, the proposed \((c, k)\)-anonymization algorithm execution time comparatively increases because of satisfying privacy equations for all the available attributes; however, the algorithm execution time is still small and acceptable.

6.4. Discussion. The proposed algorithm has been analysed through utility loss and privacy loss metrics. The main goal of the algorithm is to prevent attribute disclosures using some background knowledge. The algorithm achieves this goal. The main priority of the algorithm is to prevent from the fcorr attack. For this purpose, the algorithm creates a classification strategy in the form of \(\text{CtgT}^{\mathcal{D}}\) to have \(c\)-diverse records in each FB and to prevent attribute disclosures. In the defensive position, the algorithm focuses on reducing the \(\mathcal{D}\) between FBs using \(c_f\) in order to reduce the number of possible attacks. In the next phase, we enforce each FB for
the fulfilment of equation (4) and in the worst case equation (5) which completely avoids the risks of fcorr attack. Partitioning of attributes based on weights and enforcement for conditions reduces the privacy loss, and creating closest $k$-anonymous QI class also results in low information loss. The evaluation parameters for privacy and utility proved that there is a minimum disclosure risk and information loss for the proposed $(c, k)$-anonymization algorithm.

7. Conclusion

In this paper, privacy for MSAs of healthcare data has been addressed. Irrespective of adopting any predefined methodology for our proposed approach similar to $(p, k)$-angelization, we proposed a novel algorithm, $(c, k)$-anonymization for privacy and utility improvement in MSAs. The proposed algorithm consists of two major steps. First, it categorizes the MSAs based on calculated weights. Second, FBs are created and refined iteratively to implement privacy. To categorize the MSAs, the weight calculation is done as in [22]. The privacy risk is reduced by implementing one-to-many linking which disassociate the buckets in GT and ST. The one-to-many linking not only reduces the probability of adversary’s attacks but also improves the utility in GT. The major step in privacy implementation is to reduce the correlation between $\mathcal{I}$ FBs by satisfying equations (4) and (5). Such measures prevent the main cause of privacy breach, i.e., correlation between SAs. It makes the adversary unable to disclose the privacy of an intended individual. The experiment’s results show that both with respect to utility and privacy the proposed $(c, k)$-anonymization algorithm performs well as compared to the $(p, k)$-Angelization algorithm.

Preserving the same privacy, this work can be extended to 1:M microdata [28]. Another challenging future work can be combining 1:M with MSA in a dynamic data publishing [6, 16] scenario. To the best of our knowledge, for the later scenario, there exists no available literature.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.
Acknowledgments

This work was supported by the National Natural Science Foundation of China (61932005, 61601051, and 61941105), Beijing Municipal Science and Technology Project (Z181100003218005), and 111 Project of China B16006.

References

[1] L. Sweeney, “k-anonymity: a model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[2] F. Song, T. Ma, Y. Tian, and M. Al-Rodhaan, “A new method of privacy protection: random k-anonymous,” IEEE Access, vol. 7, pp. 75434–75445, 2019.

[3] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “l-diversity: privacy beyond k-anonymity,” ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, pp. 1–52, 2007.

[4] N. Li, T. Li, and S. Venkitasubramaniam, “t-closeness: privacy beyond k-anonymity and l-diversity,” in Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, pp. 106–115, Istanbul, Turkey, April 2007.

[5] J. Wang, K. Du, X. Luo, and X. Li, “Two privacy-preserving approaches for data publishing with identity reservation,” Knowledge and Information Systems, vol. 60, no. 2, pp. 1039–1080, 2019.

[6] X. Xiao and Y. Tao, “m-invariance: towards privacy preserving re-publication of dynamic datasets,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 689–700, Beijing, China, 2007.

[7] Q. Liu, H. Shen, and Y. Sang, “Privacy-preserving data publishing for multiple numerical sensitive attributes,” Tsinghua Science and Technology, vol. 20, no. 3, pp. 246–254, 2015.

[8] L. Zhang, J. Xuan, R. Si, and R. Wang, “An improved algorithm of individuation k-anonymity for multiple sensitive attributes,” Wireless Personal Communications, vol. 95, no. 3, pp. 2003–2020, 2017.

[9] R. Wang, Y. Zhu, T.-S. Chen, and C.-C. Chang, “Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness,” Journal of Computer Science and Technology, vol. 33, no. 6, pp. 1231–1242, 2018.

[10] L. E. E. Yong Ju and L. E. E. Kyung Ho, “Re-identification of medical records by optimum quasi-identifiers,” in Proceedings of the 19th International Conference on Advanced Communication Technology (ICACT), pp. 428–435, Bongpyeong, South Korea, 2017.

[11] N. Maheshwarkar, K. Pathak, and N. S. Choudhari, “k-anonymity model for multiple sensitive attributes,” International Journal of Computer Applications—IJCA, vol. 10, pp. 51–62, 2012.

[12] T. S. Gal, Z. Chen, and A. Gangopadhyay, “A privacy protection model for patient data with multiple sensitive attributes,” International Journal of Information Security and Privacy, vol. 2, no. 3, 2008.

[13] M. Guo, Z. Liu, and H.-B. Wang, “Personalized privacy preserving approaches for multiple sensitive attributes in data publishing,” DEStech Transactions on Engineering and Technology Research, no. ssme-ist, 2016.

[14] P. Usha, R. Shriram, and S. Sathishkumar, “Multiple sensitive attributes based privacy preserving data mining using k-anonymity,” International Journal of Scientific and Engineering Research, vol. 5, no. 4, pp. 122–126, 2014.

[15] F. Amiri, N. Yazdani, A. Shakery, and A. H. Chineai, “Hierarchical anonymization algorithms against background knowledge attack in data releasing,” Knowledge-Based Systems, vol. 101, pp. 71–89, 2016.

[16] S. A. Onashoga, B. A. Bamiro, A. T. Akinwale, and J. A. Oguntuase, “A dynamic privacy preserving data publishing technique for multi sensitive attributes,” Information Security Journal: A Global Perspective, vol. 26, no. 3, pp. 121–135, 2017.

[17] K. Ashoka and B. Poornima, “Enhanced utility in preserving privacy for multiple heterogeneous sensitive attributes using correlation and personal sensitivity flags,” in Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 970–976, Karnataka, India, September 2017.

[18] N. V. S. Lakshimpathi Raju, M. N. Seetaramanan, and P. Srinivasa Rao, “A novel dynamic KGc-slice publishing prototype for retaining privacy and utility of multiple sensitive attributes,” International Journal of Information Technology and Computer Science, vol. 11, no. 4, pp. 18–32, 2019.

[19] Y. Xu, T. Ma, M. Tang, and W. Tian, “A survey of privacy preserving data publishing using generalization and suppression,” Applied Mathematics & Information Sciences, vol. 8, no. 3, pp. 1103–1116, 2014.

[20] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing: a survey of recent developments,” ACM Computing Surveys (CSUR), vol. 42, no. 4, pp. 1–53, 2010.

[21] A. Majeeed, F. Ullah, and S. Lee, “Vulnerability-and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data,” Sensors, vol. 17, no. 5, pp. 1–23, 2017.

[22] A. Anjum, N. Ahmad, S. U. R. Malik, S. Zubair, and B. Shahzad, “An efficient approach for publishing microdata for multiple sensitive attributes,” The Journal of Super-computing, vol. 74, no. 10, pp. 5127–5155, 2018.

[23] Y. Tao, H. Chen, X. Xiao, S. Zhou, and D. Zhang, “Angel: enhancing the utility of generalization for privacy preserving publication,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 7, pp. 1073–1087, 2009.

[24] J. Han, F. Luo, J. Lu, and H. Peng, “SLOMS: a privacy preserving data publishing method for multiple sensitive attributes microdata,” Journal of Software, vol. 8, no. 12, pp. 3096–3104, 2013.

[25] V. S. Susan and T. Christopher, “Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes,” SpringerPlus, vol. 5, no. 1, pp. 964–984, 2016.

[26] X. Xiao and Y. Tao, “Anatomy: simple and effective privacy preservation,” in Proceedings of the 32nd International Conference on Very Large Data Bases VLDB Endowment, pp. 139–150, Seoul, South Korea, September 2006.

[27] J. Domingo-Ferrer and J. Soria-Comas, “From t-closeness to differential privacy and vice versa in data anonymization,” Knowledge-Based Systems, vol. 74, pp. 151–158, 2015.

[28] Q. Gong, J. Luoao, M. Yang, W. Ni, and X.-B. Li, “Anonymizing 1: M microdata with high utility,” Knowledge-Based Systems, vol. 115, pp. 15–26, 2016.

[29] Y. Shi, Z. Zhang, H. C. Chao, and B. Shen, “Data privacy protection based on micro aggregation with dynamic sensitive attribute updating,” Sensors (Basel), vol. 18, no. 7, p. 2307, 2018.

[30] J. Domingo-Ferrer and J. Soria-Comas, “Steered micro-aggregation: a unified primitive for anonymization of data sets and data streams,” in Proceedings of the 2017 IEEE
International Conference on Data Mining Workshops, pp. 995–1002, New Orleans, LA, USA, November 2017.

[31] Z. Li and X. Ye, "Privacy protection on multiple sensitive attributes," in Proceedings of the International Conference on Information and Communications Security, pp. 141–152, Zhengzhou, China, December 2007.

[32] T. Yi and M. Shi, "Privacy protection method for multiple sensitive attributes based on strong rule," Mathematical Problems in Engineering, vol. 2015, Article ID 464731, 14 pages, 2015.

[33] J. Liu, J. Luo, and J. Z. Huang, "Rating: privacy preservation for multiple attributes with different sensitivity requirements," in Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW) IEEE, pp. 666–673, Vancouver, Canada, December 2011.

[34] F. Luo, J. Han, J. Lu, and H. Peng, "ANGELMS: a privacy preserving data publishing framework for microdata with multiple sensitive attributes," in Proceedings of the International Conference on Information Science and Technology (ICIST), pp. 393–398, IEEE, Yangzhou, China, 2013.

[35] X.-C. Yang, Y.-Z. Wang, B. Wang, and G. Yu, "Privacy preserving approaches for multiple sensitive attributes in data publishing," Chinese Journal of Computers, vol. 31, no. 4, pp. 574–587, 2008.

[36] Y. Jing and W. Bo, "Personalized l-diversity algorithm for multiple sensitive attributes based on minimum selected degree first," Journal of Computer Research and Development (China), vol. 49, no. 12, pp. 2603–2610, 2012.

[37] Y. Ye, Y. Liu, and D. Lv, "Decomposition: privacy preservation for multiple sensitive attributes," in Database Systems for Advanced Applications, Lecture Notes in Computer Science, vol. 5463, pp. 486–490, Springer, Berlin, Germany, 2009.

[38] D. Das and D. K. Bhattacharyya, "Decomposition: improving l-diversity for multiple sensitive attributes," in Proceedings of the International Conference on Computer Science and Information Technology, pp. 403–412, Bangalore, India, 2012.

[39] T. M. Truta and B. Vinay, "Privacy protection: p-sensitive k-anonymity property," in Proceedings 22nd International Conference on Data Engineering Workshops, p. 94, IEEE, Atlanta, Georgia, April 2006.

[40] C. N. Sowmyarani and G. N. Srinivasan, "A robust privacy preserving model for data publishing," in Proceedings of the International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6, IEEE, Coimbatore, India, 2015.

[41] A. Abdalaal, M. E. Nergiz, and Y. Saygin, "Privacy-preserving publishing of opinion polls," Computers & Security, vol. 37, pp. 143–154, 2013.

[42] B. Al Bouna, C. Clifton, and Q. Malluhi, "Efficient sanitization of unsafe data correlations," in Proceedings of the Workshops of the EDBT/ICDT 2015 Joint Conference, pp. 278–285, Brussels, Belgium, March 2015.

[43] H. Wang and R. Liu, "Hiding outliers into crowd: privacy-preserving data publishing with outliers," Data & Knowledge Engineering, vol. 100, pp. 94–115, 2015.

[44] S. U. R. Malik, S. U. Khan, and S. K. Srinivasan, "Modeling and analysis of state-of-the-art VM-based cloud management platforms," IEEE Transactions on Cloud Computing, vol. 1, no. 1, p. 1, 2013.