Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8416823, 18 pages
https://doi.org/10.1155/2020/8416823

Research Article

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

Razaullah Khan,[1] **Xiaofeng Tao**,[1] **Adeel Anjum**,[2] **Haider Sajjad**,[2] **Saif ur Rehman Malik**,[3] **Abid Khan**,[2] **and Fatemeh Amiri**[4]

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

Academic Editor: Ghufran Ahmed

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.

Table 1: Original data table T.

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

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 [1] 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 fcorr attack is comparatively considered as a strong privacy attack. If the adversary is intended to disclose privacy of almost every individual, the fcorr attack iteratively can breach the privacy of the whole dataset. The privacy breach scenario is explained in Section 1.1 in detail.

1.1. Motivation. The (p, k)-Angelization [22] algorithm directly adopts the single SA approach named as angelization [23] to implement privacy for MSAs. This approach invalidates the (p, k)-Angelization for 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:

Table 2: Generalized table (GT) from the (p, k)-Angelization.

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

(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

Table 3: Sensitive batch table (SBT) from the (p, k)-Angelization.

Physician	Cancer type	Cancer treatment	Diagnosis method	Symptom	Diagnosis date	Batch ID
Daisy, Alan	Colon, Breast	Surgery, Biologic therapy	Blood test, MRI test	Back pain, Skin irritation	22/11/19, 13/09/19	1
Tom, Jack	Prostrate, Liver	Ablation, Biologic therapy	CT scan, Chest X-ray, Blood test	Weight loss, Abdominal pain, Back pain	5/08/19, 21/09/19, 29/08/19	2
Alan, Jack	Rectal, Breast	Inhibitor therapy, Surgery	Chest X-ray, Ultrasound,	Back pain, Swelling	7/8/19, 17/8/19	3
Tom, Frank	Prostrate, Rectal	Radiation, Medication	Blood test, CT scan	Testis swelling, Abdominal pain	12/08/19, 9/09/19	4

individual and can identify the complete records in data table T. Since the (p, k)-Angelization blindly follows the angelization [23] mechanism, correlating the MSAs in different buckets may result in single SA values against each SA. This is a column-wise vertical correlation between two SA fingerprint buckets (SAFBs) in SBT that has common physicians and other SA values. The intruder takes intersection of SAFB 3 with groups having common physicians and proceeds iteratively until p2 is identified. So, he takes intersection between SAFB 3 and SAFB 2 because of the common physician Jack, between SAFB 2 and SAFB 4 because of Tom, and then between SAFB 3 and SAFB 1 because of Alan. Table 4 depicts the identified SVs and hence the disclosed individuals. Although the intruder was interested to identify only p2, the privacy of p1 and p4 was also breached during the process, which implies that this process iteratively can breach the individuals in the complete dataset.

The intruder uses Table 3 (SBT) and on each step stores the values in Table 4 and finally identifies all the sensitive information related to p2. In Table 4, the values against each physician attribute are the values obtained by taking intersection between two SAFBs in Table 3 linked through common physician's names. In Table 3, Jack is common between SAFB 3 and SAFB 2, so whatever value David gets from intersection, he adds against Jack in Table 1. First, chest X-ray is common in the diagnostics method. The leftover value ultrasound for sure belongs to Alan. While in group 3, both the remaining diagnosis values cannot be assigned to Tom, as Tom may have only one value, so the intruder is not sure at this stage. In the symptoms attribute, back pain is common and is stored against Jack. Here, another symptom value "*swelling*" is definitely for Alan because there is neither physician nor symptom. Since any further intersection for cancer treatment and cancer type does not produce any value, the process is forwarded to SAFB 2 and SAFB 4 because of Tom. Similarly, for the diagnostic method, Tom had CT scan and Blood test and no value for Frank. Although there is one value for Frank, the intruder can refine this while taking Frank intersection with other SAFBs that are not in the current sample dataset. In the symptom column, abdominal pain for Tom and the lifted value in SAFB 4 is testis swelling for Frank. The weight loss and back pain symptoms in SAFB 2 cannot be assigned to either Tom or Jack because the intruder has no enough information about this yet. For cancer treatment, there is no common value while for the cancer type *prostrate* is assigned to Tom. The last intersection process is between 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 ($\mathscr{L}$) 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 $\mathscr{L}$ 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

Table 4: Privacy breach in the (p, k)-Angelization.

Physician	Diagnosis method	Symptoms	Cancer treatment	Cancer type	Privacy breached
Jack	Chest X-ray (intersection between bucket 2 and 3)	Back pain (intersection between bucket 2 and 3)	—	Rectal (lifted after Alan got the breast)	p1
Alan	Ultrasound (after Jack got the chest X-ray between bucket 2 and 3) MRI test (from general knowledge of rays type treatment since in bucket 3 only ray-type methods exist)	Swelling (lifted symptom after Jack got the back pain from intersection between bucket 2 and 3) Back pain (temporary)	Surgery (wrong, but not very important)	Breast (intersection between bucket 1 and 3)	p2
Tom	CT scan, blood test (intersection between bucket 2 and 4)	Abdominal pain (intersection between bucket 2 and 4)	—	Prostrate (intersection between bucket 2 and 4)	p4
Frank	—	Testis swelling	—	—	—
Daisy	—	—	—	Colon (after Alan got the breast)	—

Table 5: (c, k)-anonymization generalized table (GT).

Name	Gender	Age	Zipcode	Bucket ID
p2 (Lisa)	F	[21–27] (21)	[34548–37506] (34607)	1
p9 (Emily)	F	[21–27] (22)	[34548–37506] (34548)	1
p3 (Richard)	M	[21–27] (26)	[34548–37506] (37506)	2
p7 (Robert)	M	[21–27] (27)	[34548–37506] (35064)	3
p8 (Olivia)	*	[32–38] (32)	[43674–44662] (44662)	1
p6 (William)	*	[32–38] (38)	[43674–44662] (43674)	2
p4 (Dave)	*	[31–35] (31)	[33753–34549] (34549)	3
p1 (Michael)	*	[31–35] (34)	[33753–34549] (34548)	1
p5 (Kate)	*	[31–35] (35)	[33753–34549] (33753)	3

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. The idea of marginal publication was introduced in [23]. The 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, named as the (c, k)-anonymizaiton 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

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	Prostrate, 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

[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 fulfils 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 $T = \{\text{EI}, \text{QI}, S\}$ (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 $t \in T$ are explicit identifier attributes (also called identifying attribute) $\text{EI} = \{A_1^{ei}, A_2^{ei}, A_3^{ei}, \ldots, A_h^{ei}\}$, quasi identifier attributes (also called partial identifiers) $\text{QI} = \{A_1^{qi}, A_2^{qi}, A_3^{qi}, \ldots, A_d^{qi}\}$, and sensitive information attributes $S = \{A_1^s, A_2^s, A_3^s, \ldots, A_m^s\}$. QIs are the partial identifiers or personally identifiably information (PII) that can identify an individual i 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.

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: $\text{FB}_1, \text{FB}_2, \ldots, \text{FB}_m$ according to the following conditions:

Table 7: Summary of notations used.

Symbol	Description
T	Original microdata
T^*	Anonymized form of T
A_i^{ei}	Explicit identifiers in T
A_i^{qi}	Quasi-identifiers in T
A_i^s	Sensitive information in T
GT	Generalized table
ST	Sensitive table
BID	Bucket identifier
ED	External dataset
GB	Generalized bucket
SAFB	Sensitive attribute fingerprint bucket
EC	Equivalence class
HLPN	High-level petri nets
ctg$T$	Category table for SAs
nmk	Nonmembership knowledge
fcorr	Fingerprint correlation knowledge
$\varnothing$	Null or zero
χ	External factor
w_i	Weight of dependent A_i^s
ξ_{w_i}	Maximum weighted attributes
ζ_{w_i}	Minimum weighted attributes
C	Category level (diversity) of SAs in Ctg$T$
$\mathscr{L}$	Linkability or linking between SAFBs
$\mathscr{c}_f$	Linkability control factor
k	k-anonymous level
δ	Single maximum weighted attribute after χ

(i) Each FB consists of two columns BID and MSA values.

(ii) $\cup_{i=1}^{m} \mathrm{FB}_i = \mathrm{ST}$, and for any $i \neq j$, $\mathrm{FB}_i \cap \mathrm{FB}_j = \varnothing$ or >1 among linkable ($\mathscr{L}$) buckets through maximum weighted sensitive attributes (δ) (see Section 5).

(iii) Each SAFB, FB_i ($1 \leq i \leq m$) fulfils c-diversity from the category table (Table 8). The subscript i of FB_i is the BID of bucket FB_i.

Definition 3 (generalized bucket (GB)). A generalized bucket partitioning from an EC of the microdata T consists of buckets $B_1^{qi}, B_2^{qi}, B_3^{qi}, \ldots, B_n^{qi}$ such that

(i) Each B_i^{qi} is the set of tuples only with QI attributes of T and BID from FB

(ii) $\cup_{i=1}^{n} B_i^{qi} = \mathrm{GT}$, and for any $i \neq j$, $B_i^{qi} \cap B_j^{qi} = \varnothing$

(iii) Each generalized bucket B_i^{qi} ($1 \leq i \leq n$) fulfils k-anonymity principle

3.1. Adversarial Model. In literature, an adversary is the attacker, who intends to breach the privacy and has different types of knowledge known as bk. Data correlation is an important type of adversary knowledge that breaches the privacy. The data correlation can be attribute correlation that exists among two or more attributes, e.g., [42], or row correlation between two or more rows, e.g., [43]. This paper is related to row correlation and more specifically FB correlation. This work focuses on reducing the threat exposed 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., $\mathrm{FB}_i \cap \mathrm{FB}_j$, 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) $= \{A_1^{qi}, A_2^{qi}, A_3^{qi}, \ldots, A_d^{qi}, \mathrm{BID}\}$

(ii) ST (sensitive table) $= \{A_1^s, A_2^s, A_3^s, \ldots, A_m^s, \mathrm{BID}\}$

(iii) Any external dataset $\mathrm{ED} = \{\mathrm{ED}_1, \mathrm{ED}_1, \ldots, \mathrm{ED}_d,\}$ 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_i}) (see 5.1 algorithm). The ζ_{w_i} attributes do no 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 $N = \{P, T, F, \varphi, R_n, L, \text{and } M_0\}$, where P is a set of places represented by circles, T is a set of transitions represented by rectangular boxes such that $P \cup T \neq \varnothing$ and $P \cap T = \varnothing$, F is the flow relations such that $F \subseteq (P \times T) \cup (P \cup T)$, 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.

Table 8: Category table from original microdata table T.

Category ID	Physician	Cancer type	Cancer treatment	Diagnosis method	Symptom	Diagnosis date	Sensitivity level
One	Daisy, Frank	Colon, prostrate	Surgery, radiation	Blood test	Back pain, testis swelling	22/11/19, 9/09/19	Top secret
Two	Jack	Rectal, prostrate	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	Prostrate, 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

4. Critical Review for (p, k)-Angelization with fcorr Attack Identification Using Formal Modelling and Analysis

Definition 8 ((p, k)-Angelization [22]). A pair of bucket partitioning $= \{B_1, B_2, B_3, \ldots, B_n\}$ and batch partitioning $= \{C_1, C_2, C_3, \ldots, C_n\}$ 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 fcorr that causes fcorr 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 fcorr 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 fcorr 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 fcorr 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 ε 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.

$$\text{Rule 1: } R(\text{Chk Dep}) = \forall i2 \in x2,\ i3 \in x3 \mid i3\ [1] \coloneqq \text{Dep}\ (i2[2]_u,\ i2[2]_v)_{\forall i2[2]_u, i2[2]_v \in x2 \mid u \neq v}$$

$$\wedge x3' \coloneqq x3 \cup \{(i3[1])\}$$

$$\text{Rule 2: } R(\text{Wt} - \text{Cal}) = \forall i4 \in x4,\quad i5 \in x5 \mid i5[1] \coloneqq \text{calcWt}(|i4[1]|)$$

$$\wedge \max(i5[1]) \wedge \text{same}(i5[1]) > 1 \longrightarrow i5[1] + \varepsilon$$

$$\wedge i5' \coloneqq x5 \cup \{(i5[1])\}$$

$$\text{Rule 3: } R(\text{SortData}) = \forall i6 \in x6,\quad i7 \in x7 \mid i7[1] \coloneqq \text{wtSort}\ (|i6[1]|)$$

$$\wedge i7' \coloneqq x7 \cup \{(i7[1])\}$$

$$\text{Rule 4: } R(\text{MaxWt}) = \forall i8 \in x8,\quad i9 \in x9 \mid i9[1] \coloneqq \max \text{Wt}\ (i8[1]) \wedge$$

$$\wedge i9' \coloneqq x9 \cup \{(i9[1])\}$$

$$\text{Rule 5: } R(\text{Categorize}) = \forall i10 \in x10,\quad i11 \in x11 \mid i11[1] \coloneqq i10[1] \wedge i11[2] \coloneqq i1[2]$$

$$\wedge i11' \coloneqq x11 \cup \{(i11[1], i11[2])\}$$

The problem arises from rule 6 onward. The (p, k)-angelization [22] blindly follows the basic angelization [23] mechanism. According to rule 6, the data in table T based on the category table (Table 8) use angelization to create GT and SBT. In [23], the $\mathscr{L}$ between two SA buckets

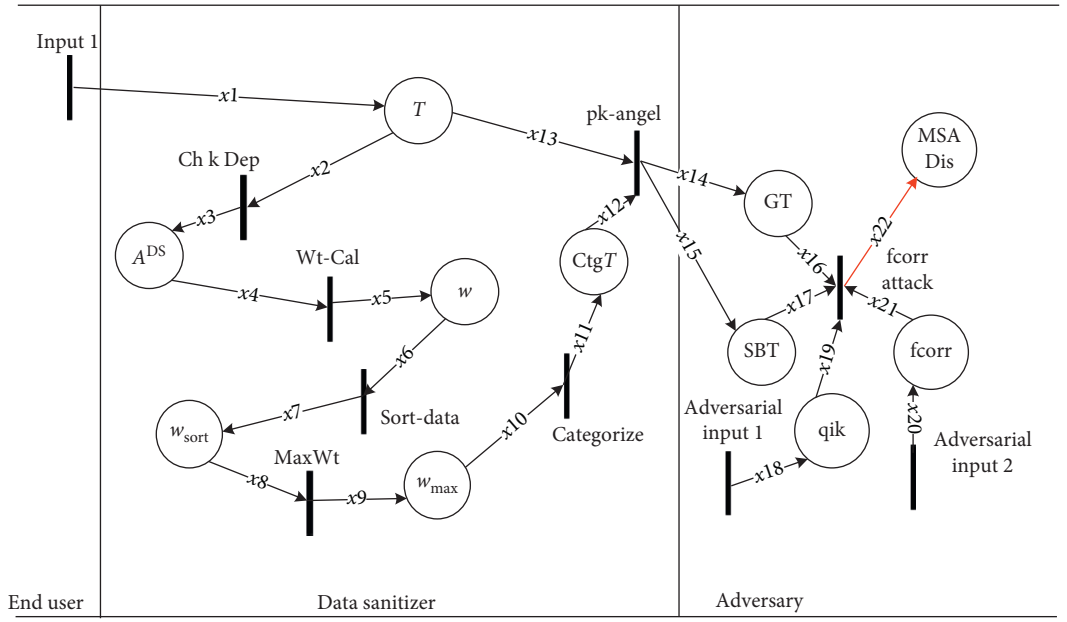

Figure 1: HLPN for (p, k)-angelization.

Table 9: Types used in HLPN for the (p, k)-angelization.

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_{sort}	Set containing sorted weight of SA in descending
w_{max}	Maximum weight after adding external factor χ
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
$\varphi(T)$	$\mathbb{P}$ (QI × MSF × PID)
$\varphi(A^{DS})$	$\mathbb{P}$ (A^{DS})
$\varphi(w)$	$\mathbb{P}$ (w)
$\varphi(w_{sort})$	$\mathbb{P}(w_{sort})$
$\varphi(w_{max})$	$\mathbb{P}(w_{max})$
φ(Ctg$T$)	$\mathbb{P}$ $(C \times MSF)$
φ(GT)	$\mathbb{P}$ (QI × BID)
φ(SBT)	$\mathbb{P}$ (MSF × BID)
φ(qik)	$\mathbb{P}$ (QI × BID)
φ(fcorr)	$\mathbb{P}$ (MSF × BID)
φ(MSA Dis)	$\mathbb{P}$ (QI × MSF × PID)

does not exist because of single SA. While handling the MSAs, the basic angelization is not applicable without proper measures. Rule 7 shows the fcorr attack that breaches the privacy of an individual. In 7, the known QI values in a specific GB in GT have exactly a single-correspondent FB in SBT for applying the fcorr attack to disclose unique values from the MSAs $\mathscr{L}$ FBs.

Rule 6: $R(\text{pk} - \text{Angel}) = \forall i12 \in x12,\ i13 \in x13, i14 \in x14, i\ 15 \in x15|$

$i14[1] \coloneqq \text{angel}(\{i12[1]\}, \{i13[1], i13[2], i13[3]\})$
$\wedge i14[2] \coloneqq \text{bucketID}(i12[1])$

$\wedge i14' \coloneqq x14 \cup \{(i14[1], i14[2])\}$

$i15[1] \coloneqq \text{angel}(\{i12[1]\}, \{i13[1], i13[2], i13[3]\}) \wedge$
$i15[2] \coloneqq \text{bucketID}(i12[1])$

$\wedge i15' \coloneqq x15 \cup \{(i15[1], i15[2])\}$

Rule 7: $R(\text{fcorr Attack}) = \forall i16 \in x16,\ i17 \in x17, i19 \in x\ 19, i21 \in x21, i22 \in x22|$

$\text{qik}i16[1], i[19[1]) \longrightarrow i22[1] \coloneqq i13[1] \wedge$

$\text{fcorrDis}(i17[1] i_{\forall i17[1]_{\mathscr{L}} \in i}, i21[1]) \longrightarrow i22[2] \coloneqq i13[2] \wedge\ i22[3] \coloneqq i13[3]$

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 [23] approach for MSAs. It leads to fcorr attack, i.e., ffcorr attack or pfcorr attack. We name the improved form of (p, k)-angelization as (c, k)-anonymization for MSAs and is describe as follows:

Definition 9 ((c, k)-anonymization). A pair of generalized bucket partitioning $= \{B_1^{qi}, B_2^{qi}, B_3^{qi}, \ldots, B_n^{qi}\}$ and sensitive attribute fingerprint bucket partitioning $= \{FB_1, FB_2, \ldots, FB_m\}$ of table T produces two tables GT and ST such that

(i) GT consists of QI attributes buckets B_i^{qi} $(1 \le i \le n)$ 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 $\cup_{i=1}^{n} B_i^{qi} = \text{GT}$, and for any $i \neq j$, $B_i^{qi} \cap B_j^{qi} = \varnothing$.

(ii) ST consists of (FB_i and BID), where BID is i ($1 \leq i \leq n$) and FB are the MSAs in buckets from table T. And $\cup_{i=1}^{m} \text{FB}_i = \text{ST}$, for any $i \neq j$, and $\text{FB}_i \cap \text{FB}_j = \varnothing$ or > 1 among linkable ($\mathscr{L}$) buckets through maximum weighted sensitive attributes (δ).

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

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

Lemma 1 (uncertainty in $\mathscr{L}$ 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 ($\text{sbp}_a \cap \text{sbp}_b > 0$ or 1, $a \neq b$) or the $\text{fcorr} \leq \zeta_{w_j} \forall\, \text{sbp}_{\mathscr{L}}$, where ζ_{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 fingerprint by k records. Thus, sbp satisfies the definition of (c, k)-diversity and uncertainty in $\mathscr{L}$ SAFBs is maximized. □

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 $\text{PriRisk}(t) = (r/n)$, 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 $\text{PriRisk}(t) \leq (1/l)$, where l represents the l-diversity [3]. Every FB has c-diverse fingerprints that correspond to at least k individuals. In our proposed (c, k)-anonymization, for $c = 1$, implies $l = 2$; for $c = 2$, implies $l \leq 4$; for $c = 3$, implies $l \leq 6$, and so on.

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, $\text{PriRisk}(t) = (r/n) = (4/42) \leq (1/4) \Longrightarrow 0.09 \leq 0.25$, which is a very low privacy risk. Similarly, for EC size 2, $r = 2$ and $n = 2 \times 4 + 2 \times 2 = 12$; so, $\text{PriRisk}(t) = (2/12) \leq (1/2) \Longrightarrow 0.16 \leq 0.5$. For the last EC size 3, $r = 3$ and $n = 3 \times 3 + 3 \times 4 = 21$; so, $\text{PriRisk}(t) = (3/21) \leq (1/3) \Longrightarrow 0.14 \leq 0.3$. 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 $\text{PriRisk}(t) = (2/4) \leq (1/2) \Longrightarrow 0.5 \leq 0.5$, 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 weights calculation for MSAs creates the category table (Ctg$T$) 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 wtCalc( ), 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:

$$w_i = \sum_{u=1}^{m} \text{Dep}\left(s_u, s_v\right)_{\forall v \in S \wedge u \neq v}, \tag{1}$$

where m are the total number of attributes dependent on attribute s_u. 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_u with all other SAs, i.e., s_v. 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.

Input:
T: Microdata table = {ID, QI, S}
χ: External Factor
Output:
GData: Generalized Table-GT
SAFB: Sensitive Table-ST
Procedure CK − ANONYMIZATION (T)
SAFB = { }, GData = { }
w_i = wtCalc (w, D, s, h)
$\forall \xi_{w_j} = (\max(w_i))_{\forall w_i \in W}$
$\forall \zeta_{w_j} = (\min(w_i))_{\forall w_i \in W}$
$\forall w_j \in \xi_{w_j} \cdot w_j = w_j + \chi(w_j)$
  v = Sort (w_j)
  $\delta = (\max(v))$
Ctg$T$ = Categorize (δ, MSA)
FB = CreateFB (Ctg$T$, T)
while FB ≠ { }
  if FB$_k$ satisfy equation (4) **then**
    SAFB = SAFB ∪ FB$_k$
  else
    SAFB RefineFB ( )
  end if
end while
GData = Generalization (T, SAFB, k)
return GData, SAFB

Algorithm 1: (c, k) − anonymization.

Input:
  s: sensitive attribute (SA)
  w: weight for a SA
  Dep: SA dependency
  h: height of dependency (no. of dependent SAs)
Output: list of weights for all SA, i.e., w_i
for each s_u and s_v
    Dep $(s_u, s_v)_{\forall u \neq v}$
end for
for check all dependent SAs
    $w_i = |\text{Dep}(s_u, s_v)|_{\forall v \in \text{Dep}(s_u, s_v)}$
end for
return w_i
end function

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

Table 11: MSAs weight calculation.

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

Categorizing sensitive attributes. Attribute occurring in the first location of the set w_j is selected as δ. 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 (δ) because of some other external factors χ. For example, physician information may be publically available on the Internet.

Creating FBs. FBs consist of MSAs with the BIDs column at the time of anonymization. Function 2, function CreateFB ( ), shows the whole process for creating c-diverse FBs. Create FBs mainly based on Ctg$T$ (as shown in Table 8). Let $A^s = \{A_1^s, A_2^s, \ldots, A_m^s\}$ and N are the total number of records in the microdata table T. To create a 2-anonymous 2-diverse FB, i.e,. $k = 2$, $l = 2$ (i.e., $c = 1$), for example, r_u and r_v are two different records in T such that $r_u \in \text{ctg}_x$ and $r_v \in \text{ctg}_y$, where $\text{ctg}_x \neq \text{ctg}_y$ and $\text{ctg}_i \in \text{Ctg}T$. The union of these two different records from different sensitive categories creates an FB, as shown in the following equation:

$$\text{FB}_k = \left(r_u \in \text{ctg}_x\right) \cup \left(r_v \in \text{ctg}_y\right), \tag{2}$$

where $u \neq v$ and $x \neq y$. Selecting records from different categories to create an FB is to implement the l-diversity principle in the form of c-diversity in our case. Privacy in data is all about creating an EC that prevents an intruder to breach any of its sensitive contents. Unlike [22], we focus on creating an FB that satisfies c-diversity and prevents against any attack from the adversary, e.g., fcorr attack. The fcorr attack is prevented by taking two measures: (i) minimizing the likability or linking ($\mathscr{L}$) between two FBs and (ii) uncorrelating the records or maximizing the uncertainty between the $\mathscr{L}$ FBs (explained in Refine FB).

Patients records (i.e., SVs) are high in number as compared to the physician attribute and both can normally be correlated or linked with one another. This $\mathscr{L}$ provides a reason for an attack. Therefore, to minimize the $\mathscr{L}$, we use the linkability control factor ($\mathscr{c}_f$), ($1 \leq \mathscr{c}_f \leq$ max count for specific δ). $\mathscr{c}_f$ minimizes the repetition of the same δ value in different FBs. The table (Table 6) published by the proposed algorithm has $\mathscr{c}_f = 2$, which means that a maximum of two records for a single physician can exist in an FB. This brings the existing δ values to minimum FBs and $\mathscr{L}$ is reduced. So, the chances of the fcorr attack on possible FB are ultimately reduced (Table 3 has 3 $\mathscr{L}$ FBs while Table 6 has only 1 $\mathscr{L}$ FB). The high value for $\mathscr{c}_f$ further reduces $\mathscr{L}$

Input:
$\text{FB} = \{\text{FB}_1, \text{FB}_2, \text{FB}_3, \ldots, \text{FB}_{\text{FB}}\}$, all FBs in the whole dataset T
r: source record to create FB, it can be r_u or r_v, $u \neq v$
N: no. of records in the actual dataset T
$\mathscr{c}_\ell$: linkability control factor
k: k-anonymity level (minimum FB size)
ctg_x: any category of SA in category Table 8
ctg_y: any category of SA in category Table 8
Output: FB having list of FB_k with k-anonymous records in each
$\text{FB} = \{\}$ $\text{FB}_k = \{\}$
while $N \neq \varnothing$
  if $\text{N} < 2k$ **then**
    $\text{FB}_k = N$
    $\text{FB} = \text{FB} \cup \text{FB}_k$
  else
    $r_{ux} = (\forall r_{u \cdot \delta_{\text{same}}} \in \text{ctg}_x) \leq \mathscr{c}_\ell$
    //if e.g. $\mathscr{c}_\ell = 2$, will select max 2
    //records having same physician
    $r_{vy} = \text{distinct } (\forall r_{v \cdot \delta_{\text{same}}} \in \text{ctg}_y) \leq \mathscr{c}_\ell$,
    //where $u \neq v$ and $x \neq y$
    $\text{FB}_k = \{r_{ux}\}$
    $\text{FB}_k = \{\text{FB}_k \cup \{r_{vy}\}\}$
    $\text{FB} = \text{FB} \cup \text{FB}_k$
    $N = N \backslash \text{FB}_k$
  end if
end while
return FB
end function

Function 2: *Function* CreateFB ($r_u, r_v, N, \mathscr{c}_\ell$, FB, FB_k, ctg_x, ctg_y).

but increases information loss, so a balance should be maintained between $\mathscr{c}_\ell$ and utility preservation.

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

$$\text{FB}_i \cap \text{FB}_j = \left\{A^s_{1i} \cap A^s_{1j},\ A^s_{2i} \cap A^s_{2j}, \ldots, A^s_{mi} \cap A^s_{mj}\right\}, \quad (3)$$

where $i \neq j$ and i, j are $\mathscr{L}$ via δ. 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 $\mathscr{L}$ FBs. The measure to prevent against the fcorr attack is no or minimum data expose 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:

$$\text{FB}_{\text{i}}^{w_i - \delta} \cap \left\{\text{FB}^{w_i - \delta} \Big|_j^n\right\} = 0 \text{ or} > 1, \quad \forall A^s, \quad (4)$$

Input:
$\text{FB} = \{\text{FB}_1, \text{FB}_2, \ldots, \text{FB}_n\}$: a set of FB that are linked via the same δ
k: minimum k-anonymity level
Output: set of refined $\text{FB} = \{\text{FB}_1, \text{FB}_2, \ldots, \text{FB}_n\}$, linked via same δ.
while $\text{FB} \neq \{\}$
  [**if** FB_k satisfies equation (4) **then** // Strict ECs or
OR if FB_k satisfies equation (5) **then**] // Relax ECs
    $\text{SAFB} = \text{SAFB} \cup \text{FB}_i$
  else
    $\text{SAFB} = \text{SAFB} \cup \{\text{FB}_i = \{\text{FB}_i \cup r_{v1} \in \text{FB}_j\}\} \wedge$
    $\text{SAFB} = \text{SAFB} \cup \{\text{FB}_n = \{\text{FB}_n \cup r_{v2} \in \text{FB}_j\}\}$ // where
      // $(r_{v1} \& r_{v2}) \in \text{FB}_j$, merge the FBs, FB_j dissolved
  end if
end while
return SAFB

Function 3: *Function* RefineFB( ).

where $i \neq j$ and FBs are $\mathscr{L}$ through δ, i.e., $\delta_{\text{FB}_i} = \delta_{\text{FB}_j}$, for example, they are $\mathscr{L}$ through the physician and n are the total number of FBs where the same physician exists. The intersection in this case ensures that fcorr should be zero or have more than one SVs in common to create uncertainty for single SA.

In case of the worst dataset, there may be some of the records that do not fit in any FB via the strict approach. A relax strategy is adopted with no breach in privacy. The percentage of record exposure from fcorr is minimized to an acceptable value. In the worst case scenario, the proposed (c, k)-anonymization maintains $\text{fcorr} \leq \zeta_{w_j}$, $\forall \text{FB}_{\mathscr{L}}$ where ζ_{w_j} are the minimum weighted attributes that have no dependency. Relax strategy is given in the following equation:

$$\text{FB}_i^{w_i - \left(\zeta_{w_i} \cup \delta\right)} \cap \left\{\text{FB}^{w_i - \left(\zeta_{w_i} \cup \delta\right)} \Big|_j^n\right\} = 0 \text{ or} > 1, \quad \forall A^s, \quad (5)$$

where $i \neq j$ and FBs are $\mathscr{L}$ through δ, i.e.,$\delta_{\text{EC}_i} = \delta_{\text{FB}_j}$, for example, they are $\mathscr{L}$ through physician and n are the total number of FBs where the same physician exists. According to equation (5), only $\text{fcorr} \leq \zeta_{w_j}$, $\forall \text{FB}_{\mathscr{L}}$ is acceptable which is a percentage of information leakage but not a privacy breach because of not dependent attributes and hence impossible for an intruder to link with other SA or QI of a specific patient record. The working of RefineFB() function for equation (5) is the same as it is for equation (4).

Generalization. Function 4 deals with QI attributes with BIDs that correspond to unique FBs in ST. Initially, the records are sorted by QIs. Then, every individual record is generalized to achieve k-anonymous EC. For generalizing the tuple $t \in N$, the t $\text{QI} = \{x_1, x_2, \ldots, x_n\}$ is generalized to $\text{QI}' = ([y_1 - z_1], [y_2 - z_2], \ldots, [y_n - z_n])$, where $y_i \leq x_i \leq z_i$ and y_i, z_i are the close boundaries for x_i. The Generalization() function in Function 4 shows the generalization process.

Input:
T: QI attributes from original microdata T
k: k anonymity level
BID: Bucket identifier from SAFBs
Output: GData
GData = { }
$N = |T|$ //BID is received from SAFBs
sort N w.r.t QIs
  while $N \neq \varnothing$
    if $N < 2k$ **then**
      GData = N
    else
      GData = GData $\cup$ gen (N) // using BID
  each GB has mostly both real and likely linking
    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 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 Ctg$T$ 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., $\mathbb{P}$ (QI $\times$ MSF $\times$ PID)), and based on category id C the categorized MSA are accommodated into different c-diverse FBs.

Rule 8 $R(\text{Crt} - \text{FB}) = \forall i12 \in x12, i13 \in x13, i14 \in x\ 14|$
$i14[1] \coloneqq \text{createFB}(\{i13[1]\}, \{i12[1], i12[2], i12[3]\})$
$\wedge i14[2] \coloneqq \text{bucketID}(i13[2])$
$\wedge i14' \coloneqq x14 \cup \{(i14[1], i14[2])\}$

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 $\mathscr{L}$ FBs. 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: $R(\text{Chk} - \text{Eqs}) = \forall i15 \in x15, i17 \in x17, i18\ \in x18|$
satisfy equation $(i17[1] i_{\forall i17[1]_{\mathscr{L}} \in i}) = i15[1] \longrightarrow$
$i18[1] \coloneqq \text{TRUE} \wedge i18' \coloneqq x18 \cup \{(i18[1])\}$
$\vee$satisfy equation $(\text{i17}[1] j_{\forall i17[1]_{\mathscr{L}} \in j}) \neq i15[1] \longrightarrow i18[1]$
$\coloneqq \text{FALSE} \wedge i18' \coloneqq x18 \cup \{(i18[1])\}$

Rule 10: $R(\text{Pick FB}) = \forall i19 \in x19, i20 \in x20, i21\ \in x21|$
$i20 = \text{TRUE} \longrightarrow i21[1] \coloneqq \text{storeAll}(i19[1] k_{\forall i19}$
$[1] \in k) \wedge i21[2] \coloneqq i19\ [2]$
$\wedge i21' \coloneqq x21 \cup \{(i21[1], i21[2])\}$

Rule 11: $R(\text{Call RFB}) = \forall i22 \in x22, i23 \in x23, i24\ \in x24|$
$i22 = \text{FALSE} \longrightarrow\ i24[1] \coloneqq \text{storeAll}(i23[1] j_{\forall i23[1] \in j}) \wedge$
$i24[2] \coloneqq i23[2]$
$\wedge i24' \coloneqq x24 \cup \{(i24[1], i24[2])\}$

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 SAFBr. 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: $R(\text{Crt RFB}) = \forall i25 \in x25, i26 \in x26|$
$i26[1] \coloneqq \text{refineFB}(i25[1] k_{\forall i25[1] \in k}) \wedge i26[2] \coloneqq i25$
$[2] \wedge i26' \coloneqq x26 \cup \{(i26[1], i26[2])\}$

Rule 13: $R(\text{Crt ST}) = \forall i27 \in x27, i28 \in x28, i29\ \in x29|$
$i29[1] \coloneqq \text{combine}(i27[1], i28[1]) \wedge i29[2] \coloneqq$
$\text{combine}(i27[2], i28[2])$
$\wedge i29' \coloneqq x29 \cup \{(i29[1], i29[2])\}$

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: $R(\text{Gen}) = \forall i30 \in x30, i31 \in x31, i32 \in x32,$
$i33 \in x33|$
$i33[1] \coloneqq \text{generalize}(\{i31[1]\}, \{i30[1]\})$
$\wedge i33[2] \coloneqq i32[2]$
$\wedge i33' \coloneqq x33 \cup \{(i33[1], i33[2])\}$

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 MSAs 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: $R(\text{fcorr Attack}) = \forall i34 \in x34, i35 \in x35, i37 \in$
$x37, i38 \in x38|$
$\text{fcorrDis}(i34[1], i35[1]) \longrightarrow (\{i34[1], i35[1]\} \cup$
$i37[2]) \neq i12[2] \wedge i12[3]$
$(i38[1] \cup i38[2]) = \varnothing$

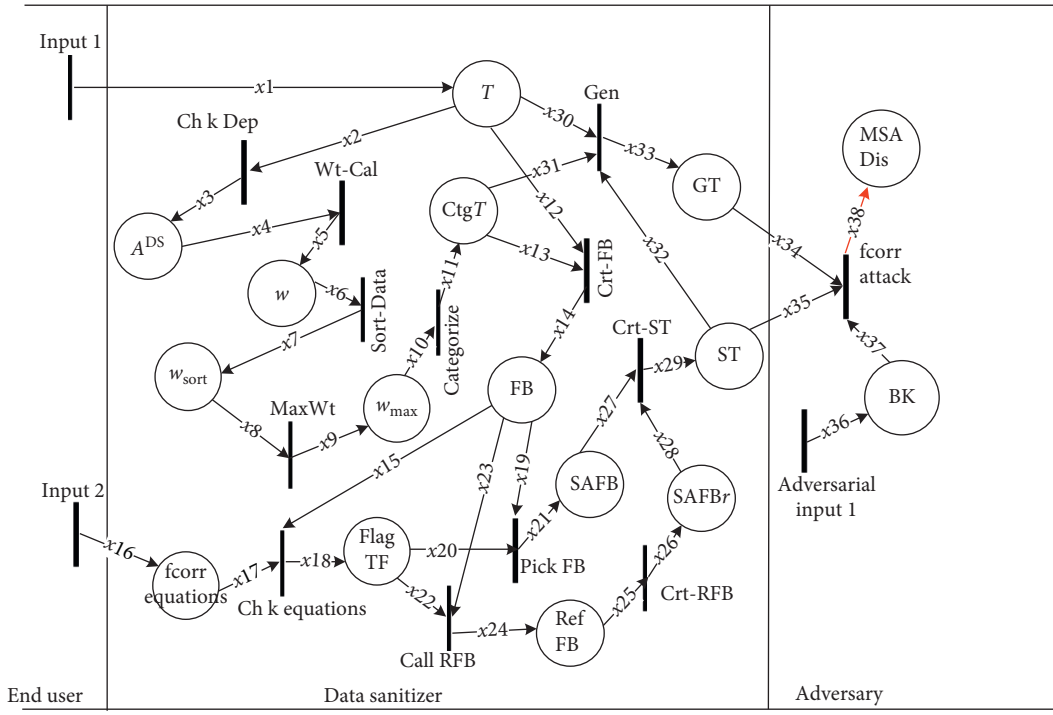

Figure 2: HLPN for the (c, k)-anonymization algorithm.

Table 12: Types used in HLPN for (c, k)-anonymization.

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_{sort}	An integer set containing sorted weight of SA in descending
w_{max}	Maximum weight after adding external factor χ
C	Category id for MSA
BID	Bucket id used to connect GT and SBT
QI	An integer and string type quasi-identifiers for i^{th} end user
MSF	Multiple sensitive fingerprint string values for i^{th} end user
MSF_I	Initial MSF string values in a bucket for i^{th} end user
MSF_N	Not correlated MSF string values in a bucket for i^{th} end user
MSF_C	Correlated MSF string values in a bucket for i^{th} end user
MSF_R	Refined MSF string values in a bucket for i^{th} end user

Table 13: Mapping of data types on places.

Places	Mapping
$\varphi(T)$	$\mathbb{P}(QI \times MSF \times PID)$
$\varphi(A^{DS})$	$\mathbb{P}(A^{DS})$
$\varphi(w)$	$\mathbb{P}(w)$
$\varphi(w_{sort})$	$\mathbb{P}(w_{sort})$
$\varphi(w_{max})$	$\mathbb{P}(w_{max})$
$\varphi(CtgT)$	$\mathbb{P}(C \times MSF)$
$\varphi(FB)$	$\mathbb{P}\ MSF_I \times BID)$
$\varphi(FlagTF)$	$\mathbb{P}(condition)$
$\varphi(SAFB)$	$\mathbb{P}(MSF_N \times BID)$
$\varphi(RefFB)$	$\mathbb{P}(MSF_C \times BID)$
$\varphi(SAFBr)$	$\mathbb{P}(MSF_R \times BID)$
$\varphi(GT)$	$\mathbb{P}(QI \times BID)$
$\varphi(ST)$	$\mathbb{P}(MSF \times BID)$
$\varphi(BK)$	$\mathbb{P}(QI \times MSF_C \times PID)$
$\varphi(MSA\ Dis)$	$\mathbb{P}(QI \times MSF \times PID)$

6. Experimental Analysis

In this section, we evaluate our proposed anonymization algorithm (c, k)-anonymization to compare the performance with (p, k)-Angelization. Both the algorithms are implemented in JAVA language on a machine having Windows 10 operating system with 4 GB RAM and Intel Core i5 2.39 GHz processor. The values plotted for the (p, k)-angelization algorithm have been obtained from the algorithm's program 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

Table 14: Attributes used in the experiments.

S. no	Attributes		Description	No. of distinct values
1	QI	Age	Age in years (range values)	48
2		Gender	Male or female	2
3	MSA	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

publisher only evaluates the similarity between the original and anonymized data. The lower values in utility loss and privacy loss reflect the effectiveness of the developed algorithm. We measure the utility loss using the normalized certainty penalty (NCP) [22], query accuracy [22], and privacy loss by calculating the vulnerable records. Also, both algorithms execution time are analysed and a discussion is provided at the end.

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 $T = \{Q_1, Q_1, \ldots, Q_q\}$ are QI. The information loss for a single QI attribute is $\mathrm{NCP}_{QI_i}(t) = (z_i - y)_i / |QI_i|$, where $y_i \le x_i \le z_i$, x_i is the actual QI value from T, and $|\mathrm{QI}_i|$ is the domain range on QI_i, i.e., max $\{t.\mathrm{QI}_i\}$–min $\{t.\mathrm{QI}_i\}$. 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:

$$\mathrm{NCP}\left(T^*\right) = \sum_{t \in T^*} \sum_{i=1}^{q} w_i \cdot \mathrm{NCP}_{Q_i}(t), \tag{6}$$

where $\mathrm{NCP}(t) = \sum_{i=1}^{q} w_i \cdot \mathrm{NCP}_{Q_i}(t)$ 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)-anonymization 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

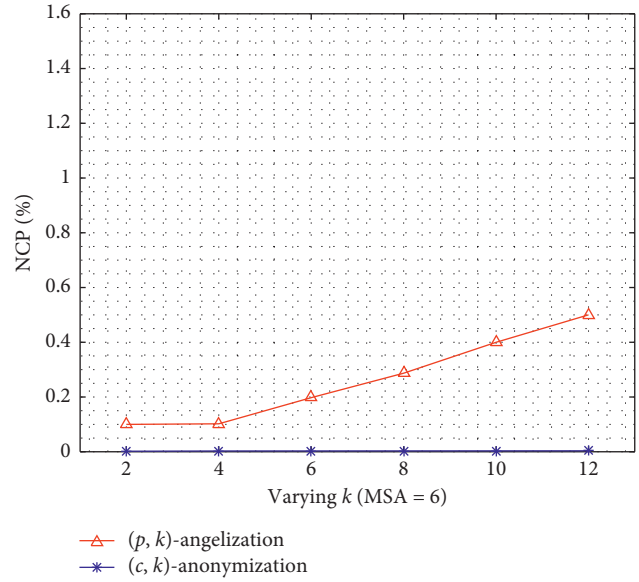

Figure 3: Normalized certainty penalty.

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:

$$\begin{aligned} &\text{Select COUNT}(*) \text{ from } T^* \text{ where } A_1^{QI} \in \text{Domain} \\ &\left(A_1^{QI}\right) \text{AND } \ldots \text{ AND } A_q^{QI} \in \text{Domain}\left(A_q^{QI}\right). \end{aligned} \tag{7}$$

Published table T^* has q as $A^{QI}s$, i.e., $A_i^{QI}, A_2^{QI}, \ldots, A_q^{QI}$. The domain (A_i^{QI}) size depends on query selectivity (θ) 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

loss, the query error in equation (8) analyses the error between the COUNT queries executed on the published and original datasets.

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

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., θ 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 record 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 $\mathscr{L}$ 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 $\mathscr{L}$ 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 $\mathscr{L}$ 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 $\mathscr{L}$ 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.

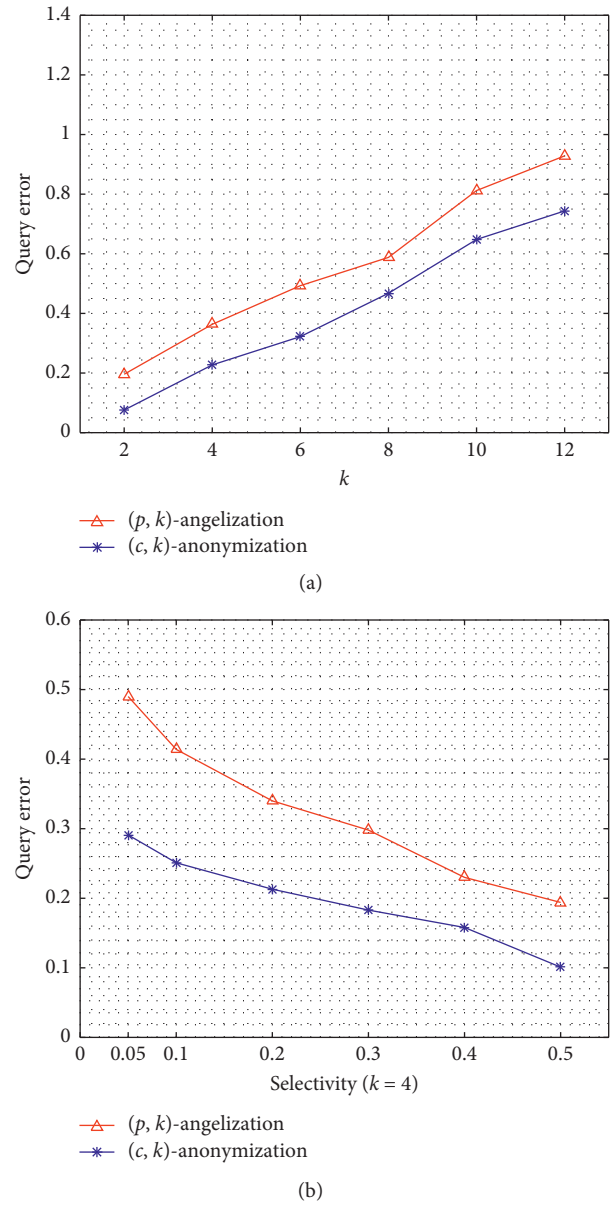

Figure 4: (a) Query error w.r.t varying k size; (b) query error w.r.t varying selectivity.

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 Ctg$T$ to have c-diverse records in each FB and to prevent attribute disclosures. In the defensive position, the algorithm focuses on reducing the $\mathscr{L}$ between FBs using $\mathscr{C}_f$ in order to reduce the number of possible attacks. In the next phase, we enforce each FB for

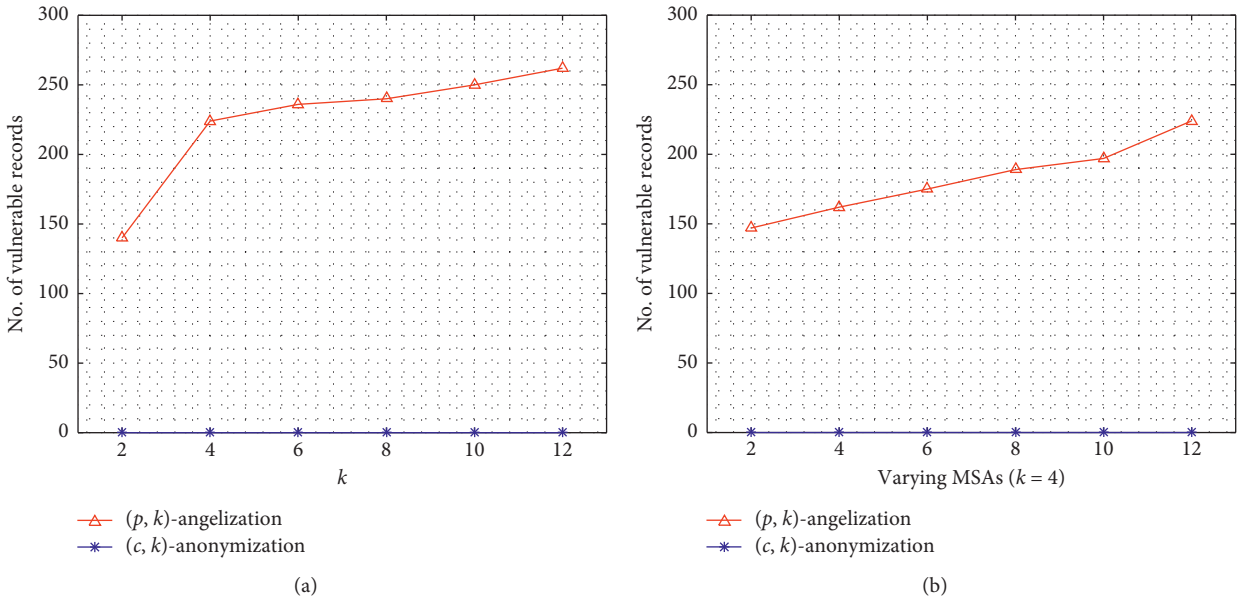

Figure 5: Privacy loss (a) with varying k size and (b) with varying MSA.

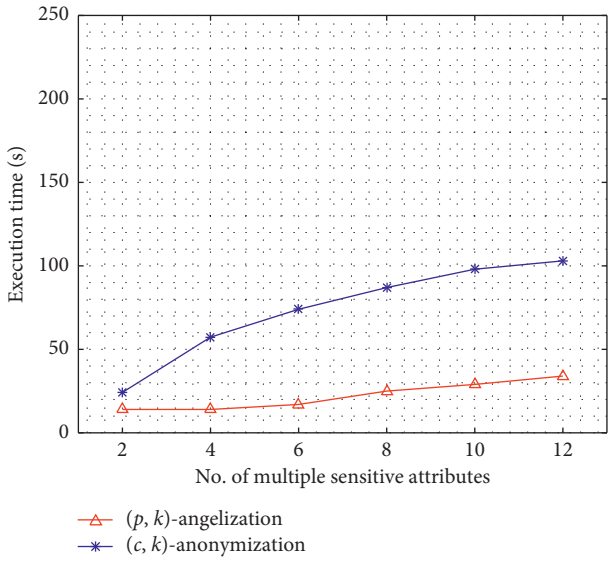

Figure 6: Algorithm execution time.

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 $\mathscr{L}$ 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. Venkatasubramanian, "*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. 51–56, 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. Chinaei, "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. Lakshmipathi Raju, M. N. Seetaramanath, and P. Srinivasa Rao, "A novel dynamic KCi-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. Majeed, 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 Supercomputing*, 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. Gonga, J. Luoa, 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 microaggregation: 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.