*electronics*

Article

# $\theta$-Sensitive $k$-Anonymity: An Anonymization Model for IoT based Electronic Health Records

**Razaullah Khan [1], Xiaofeng Tao [1,*], Adeel Anjum [2], Tehsin Kanwal [2], Saif ur Rehman Malik [3], Abid Khan [4], Waheed ur Rehman [5] and Carsten Maple [6]**

[1] National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan
[3] Cybernetica AS Estonia, Tallinn 13412, Estonia
[4] Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK
[5] Department of Computer Science, University of Peshawar, Peshawar 25120, Pakistan
[6] Warwick Manufacturing Group, The University of Warwick, Coventry CV4 7AL, UK
* Correspondence: taoxf@bupt.edu.cn; Tel.: +86-106-228-3254

Received: 28 March 2020; Accepted: 20 April 2020; Published: 26 April 2020

**Abstract:** The Internet of Things (IoT) is an exponentially growing emerging technology, which is implemented in the digitization of Electronic Health Records (EHR). The application of IoT is used to collect the patient's data and the data holders and then to publish these data. However, the data collected through the IoT-based devices are vulnerable to information leakage and are a potential privacy threat. Therefore, there is a need to implement privacy protection methods to prevent individual record identification in EHR. Significant research contributions exist e.g., $p^+$-sensitive $k$-anonymity and balanced $p^+$-sensitive $k$-anonymity for implementing privacy protection in EHR. However, these models have certain privacy vulnerabilities, which are identified in this paper with two new types of attack: the *sensitive variance attack and categorical similarity attack*. A mitigation solution, the $\theta$-sensitive $k$-anonymity privacy model, is proposed to prevent the mentioned attacks. The proposed model works effectively for all $k$-anonymous size groups and can prevent sensitive variance, categorical similarity, and homogeneity attacks by creating more diverse $k$-anonymous groups. Furthermore, we formally modeled and analyzed the base and the proposed privacy models to show the invalidation of the base and applicability of the proposed work. Experiments show that our proposed model outperforms the others in terms of privacy security (14.64%).

**Keywords:** Internet of Things; big data; electronic health records; $k$-anonymity; privacy; security

## 1. Introduction

The current highly-connected technological society generates a huge amount of digital data—termed Big Data, collected through internet-enabled devices, termed the Internet of Things (IoT) [1]. Billions of these IoT devices sense and collect the data e.g., the patient's Electronic Health Records (EHR) [1–4]. The collected data are then shared with corporate or government bodies for research and policymaking. However, the privacy of the individual records is an important goal when sharing data that is collected through the IoT enabled devices [1–6]. This is because these data contain names or some unique identification (explicit identifiers—$A^{ei}$), such as age, gender, zip code (quasi-identifiers—$A^{qi}$), and some health-related private information (sensitive attributes—$A^{s}$) [7–12]. To preserve privacy, eliminating the $A^{ei}$ before sharing or publishing the data is not enough [11]. For an attacker or an adversary, the quasi-identifiers (QIs) are the partial identifiers that can be used to link to some externally available data e.g., voting or census data, to identify an individual $A^{s}$, known as a linking attack [10–12].

To implement data privacy, a lot of cryptographic techniques [13,14] have been proposed. However, these techniques have high computational overheads. Another simple approach is data anonymization. Data anonymization is about concealing an individual's identity in a small crowd of records before data publishing. The publishing of such anonymized records are known as Privacy Preserving Data Publishing (PPDP) [11]. A plethora of PPDP methods have been proposed [7–12,15–19]. These techniques are broadly classified into:

- *Identity disclosure prevention*: Generalizing [7–9] the QI values of a group of records from more specific values to less specific values e.g., $k$-anonymity [7,8], where every record should be indistinguishable from at least $k$-1 other records. An individual having probability higher than 1/$k$ cannot be re-identified by an intruder/attacker.
- *Attribute disclosure prevention*: Preventing to reveal private information ($A^s$ information) about an individual. Examples are $l$-diversity [15] and $t$-closeness [16] privacy models.

In this paper, a variance-based privacy model is proposed to prevent *attribute disclosure* risk. For sensitive attribute privacy, the $p$+-sensitive $k$-anonymity, ($p$, $\alpha$)-sensitive $k$-anonymity [17] privacy model is a state-of-the-art privacy model where the sensitive values are categorized into four categories. For creating a $k$-anonymous group of records called an equivalence class (EC), a $l$-diversity [15] is applied. However, two new possible attacks are applied: sensitive variance attack and categorical similarity attack. These attacks breach the privacy of the $p$+-sensitive $k$-anonymity and ($p$, $\alpha$)-sensitive $k$-anonymity [17] algorithm, due to the $A^s$ values from the single sensitive category or a low diversity at $A^s$ category level. The proposed mitigation solution: the $\theta$-Sensitive $k$-anonymity privacy model, is a numerical measure of privacy strength for thwarting the attribute disclosure risk. The proposed approach also appends small amounts of noise tuple(s) to increase the variability in an EC, if needed. To minimize the utility loss, the proposed algorithm uses a bottom-up generalization (i.e., the local recoding mechanism [18]) because it minimally distorts the data compared to global recoding techniques [12]. The following section presents the motivation of our work.

### 1.1. Motivation

Broadly examining the PPDP models for preventing attribute disclosure risk [11,15–18], it was concluded that the worthiness of each model exists in the diversity of an EC where the sensitive values belongs to different categories. Such variability of $A^s$ values creates a diverse EC. Different privacy models employ different techniques to achieve variability in $k$-anonymous ECs. The repeated frequencies of same sensitive values are the only obstruct in achieving the required diversity in an EC. The privacy models in [17] and [18] provide a meaningful approach in dealing with the attribute disclosure problem, however the following limitations have been observed.

- $p^+$-sensitive $k$-anonymity and ($p$, $\alpha$)-sensitive $k$-anonymity [17]: This model is a modified version of the $p$-sensitive $k$-anonymity [19], for preventing a similarity attack. However, the $p^+$-sensitive $k$-anonymity and ($p$, $\alpha$)-sensitive $k$-anonymity models have zero diversity at the $A^s$ category level, which may lead to a categorical similarity attack. A more powerful possible attack by an adversary is the sensitive variance attack, due to the low variability at $A^s$ category level. With an upsurge in the adversary's knowledge (background knowledge—BK) the privacy level can be breached, which may cause attribute disclosures. The proposed $\theta$-sensitive $k$-anonymity privacy model provides a privacy solution to prevent all such attacks.
- Balanced $p^+$-sensitive $k$-anonymity and ($p$, $\alpha$)-sensitive $k$-anonymity [18]: This model is an enhanced version of $p^+$-sensitive $k$-anonymity model. It balances the categorical level sensitive attributes in each EC. However, it still has low diversity at the $A^s$ category level and works only for more than three $k$-anonymous size ECs.

To solve the problems of homogeneity, categorical similarity, and sensitive variance attacks in the $p^+$-sensitive $k$-anonymity and ($p$, $\alpha$)-sensitive $k$-anonymity model [17], we propose the $\theta$-Sensitive

$k$-anonymity privacy model in this paper. The categorical level similarity and small EC size problems in the balanced $p^+$-sensitive $k$-anonymity and $(p, \alpha)$-sensitive $k$-anonymity model [18] are also addressed by achieving a more balanced and diverse EC even at the category level and its execution on small $k$ size EC, i.e., $k = 2$.

### *1.2. Contributions*

The proposed $\theta$-sensitive $k$-anonymity privacy model multiplies variance ($\sigma^2$) of a fully diverse EC with an observed value (observation 1) which produces a threshold value $\theta$. The $\theta$ value ensures prevention against attribute disclosure in an EC which collectively results in the privacy of the given dataset.

The contributions of this paper are as follows:

- A new $\theta$-sensitive $k$-anonymity privacy model is proposed where privacy in an EC is achieved through a threshold value, i.e., $\theta$. The $\theta$ value for an EC is obtained by multiplying variance and an observation value. The variance-based diversity in an EC prevents the sensitive variance attack, which automatically prevents the categorical similarity attack. In the proposed model, the $A^s$ values checking is not only performed with next ECs, but a cross check is also performed during the last EC. If the required privacy is not achievable with the existing $A^s$ values, then a noise is added for the required diversity.
- We formally modeled and analyzed the base model in [17] and the proposed $\theta$-sensitive $k$-anonymity privacy model using High Level Petri-Nets (HLPN).
- Based on the above points, simulation results show that our proposed $\theta$-sensitive $k$-anonymity model has only 0.002679% higher privacy leakage than its counterpart $p^+$-sensitive $k$-anonymity model which has 14.65% higher privacy leakage with the base line privacy.

Paper Organization. The remainder of the paper is organized as follows. Section 2 explains related work. Preliminaries are discussed in Section 3. The considered attacks and problem statement in $p^+$-sensitive $k$-anonymity along with its formal analysis are presented in Section 4. Section 5 discusses the proposed $\theta$-sensitive $k$-anonymity model and its formal analysis. In Section 6, the experiments and evaluations are provided. Section 7 concludes the paper.

## 2. Related Work

In this section, the literature related to the proposed privacy model is studied from various aspects. The data collected through the various IoT enabled devices [1–6,20,21] must be anonymized before publishing because of the private information contained in it. Anonymized data are published for the sake of its maximum utility without disclosing the private information of an individual. For anonymization, the privacy models can be broadly classified into semantic [22,23] or syntactic [7–19] approaches. The semantic privacy models add a random amount of noise for preserving privacy, e.g., differential privacy models [22,23]. In differential privacy, the deletion or addition of an individual's record or noise does not affect the data analysis results while preserving the privacy. Syntactic privacy models create a $k$-indistinguishable [7] ECs. In syntactic privacy, two main privacy disclosure risks are: identity disclosure [7–10,12] and attribute disclosure [11,15–18]. The $k$-anonymity [7,8] is an example of preventing identity disclosure that generalizes a set of records with respect to QIs. These $k$-anonymous records are indistinguishable from other $k$-1 records in a dataset. However, $k$-anonymity lacks the ability to provide attribute level protection. Attribute disclosure releases the value of confidential attributes corresponding to an identified individual record. Although in $l$-diversity [15], $l$ distinct groups for the $A^s$ in an EC are required. However, the skewness and similarity attacks can breach the privacy because $l$-well sensitive attribute groups are not always possible over the existing $A^s$s. Similarly for $t$-closeness [16], the threshold for $A^s$ and its distance distribution in an EC has low data utility, and the earth mover distance (EMD) is not an efficient prevention for attribute linkage [24,25].

In [26] by Torra, identity and attribute disclosure were both addressed. Jose et al. [27] proposed an adaptive two-step iterative anonymization approach. A privacy leakage for an attribute linkage attack was possible because of having numerous versions. An extended $k$-anonymity model was proposed by Rahimi et al. [28] to protect identity and attribute information. However, a BK attack is possible because the publisher is unaware of the adversary's knowledge. The $k$-join-anonymity model proposed by Sowmiyaa et al. [29] was the same as $k$-anonymity, which focuses only on identity disclosure risk. The $(\alpha, k)$-anonymity model proposed by Wong et al. [30], used a global recoding technique, which has a high utility loss and, due to table linkage attack, it was susceptible to the disclosure of attributes.

The $(k, e)$-anonymization model proposed by Zhang et al. [31] publishes separate tables, consisting of $A^s$ and QI to reduce the relationship between them, and where instead of generalization, a permutation-based approach has been adopted. Although in aggregated search, not using QI-generalization is recommended for accuracy improvement. However, a probabilistic attack is possible over the $A^s$ due to the one-time publication of the microdata. The $(\epsilon, m)$-anonymity model [32] deals with the numeric $A^s$, however it is limited to work for categorical $A^s$. Xiao et al. [33] worked on personalized anonymity that uses a greedy personalized generalization approach. This model de-associated $A^s$ and QI instead of modifying the association between them.

In Reference [19], the $p$-sensitive $k$-anonymity found the closest neighbor. This model was then improved by Sun at al. [17] with a top-down specialization. The generated anonymized datasets should be from at least $p$ distinct $A^s$ values categories for each EC. However, the developed algorithm in [17] is vulnerable to privacy leakage from sensitive variance, categorical similarity, and homogeneity attacks. In this paper, these privacy limitations were mitigated using the proposed $\theta$–sensitive $k$-anonymity algorithm. The proposed privacy model is a syntactic privacy model for preventing attribute disclosure risk, which adds a fixed amount of noise to create $k$-anonymous ECs.

## 3. Preliminaries

Let an original Microdata Table (MT ) $= \{\mathrm{EI}, \mathrm{QI}, \mathrm{S}\}$ (i.e., Table 1a) be the private static data (i.e., one-time release) for a publisher to publish. The $t \in$ MT is a tuple that belongs to an individual i, such that EI = $\{A^{ei}_1, A^{ei}_2, A^{ei}_3 \dots A^{ei}_h\}$, QI = $\{A^{qi}_1, A^{qi}_2, A^{qi}_3 \dots A^{qi}_m\}$, and S = $\{A^s\}$ (this work considers only single $A^s$). The $k$-anonymized data essentially consists of $A^{qi}$ and $A^s$, while $A^{ei}$s are removed. This is because an adversary can link the $A^{qi}$ with some external information (e.g., voter or census data) to perform a record linkage attack (i.e., identity disclosure) [34]. However, the $k$-anonymous $A^{qi}$ values prevent the record against the record linkage attack in an EC. For example, consider some common diseases in a 2-anonymous (Table 1b) obtained from the original microdata Table 1a. Table 2 summarizes the notations used in this paper.

**Table 1. a**. Original microdata. **b**. 2-Anonymous microdata.

| **ID** | **Name** | **Age** | **Zip Code** | **Country** | **Disease** |
|---|---|---|---|---|---|
| 1 | JULIAN | 34 | 14247 | USA | HIV |
| 2 | KALEEM | 40 | 14208 | Pakistan | HIV |
| 3 | JOHANNA | 26 | 14205 | USA | Cancer |
| 4 | MICHAEL | 25 | 14242 | Canada | Cancer |
| 5 | JUDITH | 40 | 14054 | USA | Hepatitis |
| 6 | EVA | 48 | 13073 | Japan | Phthisis |
| 7 | HARIS | 45 | 14066 | Pakistan | Asthma |
| 8 | PAUL | 40 | 14063 | USA | Obesity |
| 9 | YIN LI | 40 | 14243 | China | Flu |
| 10 | BEVERLY | 37 | 14203 | Canada | Flu |
| 11 | DENISE | 36 | 14204 | Canada | Flu |
| 12 | JANETTE | 35 | 14247 | USA | Indigestion |

(**a**)

| **ID** | **Age** | **Zip Code** | **Country** | **Disease** |
|---|---|---|---|---|
| 1 | 34–40 | 14208-14247 | ** | HIV |
| 2 | 34–40 | 14208-14247 | ** | HIV |
| 3 | 25–26 | 14205-14242 | America | Cancer |
| 4 | 25–26 | 14205-14242 | America | Cancer |
| 5 | >= 40 | 14054-14063 | America | Hepatitis |
| 6 | >= 40 | 14054-14063 | America | Obesity |
| 7 | >= 40 | 13073-14066 | Asia | Asthma |
| 8 | >= 40 | 13073-14066 | Asia | Phthisis |
| 9 | 35–40 | 14243-14247 | ** | Flu |
| 10 | 35–40 | 14243-14247 | ** | Indigestion |
| 11 | 36–37 | 14203-14204 | America | Flu |
| 12 | 36–37 | 14203-14204 | America | Flu |

(**b**)

**Definition 1.** *k-anonymity [7,8]: Relation R having $A^{qi}$ over the schema R($A_1$,$A_2$, . . . , $A_n$) in a masked microdata table T′ is said to be k-anonymous if and only if, for any combination $A_i^{qi}$×* t($A_{in}^{qi}$) *values from start to end, is greater than or equal to k in R.*

$$\text{iff } |\forall\{A_i^{qi} \times t(A_{in}^{qi})\}| \in T' \geq k$$

*where k is the anonymity level (as shown in Table 1b). The k-anonymity model blends the k records into at least a k-1 crowd but it does not impose any restrictions on the algorithm to sufficiently protect the individuals. Consequently, the probability of linking a victim to a specific record through $A^{qi}$s is at most 1/k.*

**Definition 2.** *l-Diversity [15]: A QI block in a masked microdata table T′ having m QI-blocks $QI_j$* $(1 \leq j \leq m)$ *is l-diverse, if it contains more than or equal to l well significant $A^s$ values. In an l-diverse modified microdata table T′, every QI block is l-diverse.*

$$\text{iff } |\forall\{A_i^{qi} \times A_i^{s}\}| \in T' \geq l$$

**Definition 3.** *t-closeness [16]: An EC is considered as t-closed if the distance between the distribution of the sensitive data in a class and the distribution of sensitive data in the whole table is equal to or less than threshold t. If every EC is t-closed, the whole table is t-closed. To calculate the distance while studying the transportation problem, researchers have explored some methods [33,35]. However, most of them focused on the Earth Mover Distance (EMD) method [15,36]. The EMD(P, Q) measures the minimum cost for transforming one distribution P to another distribution Q. It depends on the amount and distance of mass moved.*

**Table 2.** Summary of notations used.

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| MT | Microdata Table | $\mathrm{A}_{\mathrm{i}}^{\mathrm{qi}}$ | Quasi identifier for $\mathrm{i}^{\mathrm{th}}$ end user |
| MMT | Micro Mask Table | $\mathrm{A}^{\mathrm{s}}$ | Sensitive Attributes |
| A | Attributes in MT | $\mathrm{A}^{\mathrm{id}}$ | Identifier Attribute |
| PD | Published Data | $\mathrm{A}_{\mathrm{ECc}}^{\mathrm{s}}$ | Sensitive value in an $\mathrm{EC}_{\mathrm{c}}$ |
| ECs | Set of Equivalence classes | $\mathrm{A}_{\mathrm{ECn}}^{\mathrm{s}}$ | Sensitive value in an $\mathrm{EC}_{\mathrm{n}}$ |
| $\mathrm{EC}_{\mathrm{i}}$ | *k*-anonymous group of tuples with the combination of $\mathrm{A}_{\mathrm{i}}^{\mathrm{qi}}$ and $\mathrm{A}^{\mathrm{s}}$ | $\mathrm{A}_{\mathrm{ECn-1}}^{\mathrm{s}}$ | Sensitive value in an $\mathrm{EC}_{\mathrm{n-1}}$ |
| $\mathrm{EC}_{\mathrm{c}}$ | Equivalence Class current | $\mathrm{A}_{\mathrm{ECb}}^{\mathrm{s}}$ | Sensitive value in an $\mathrm{EC}_{\mathrm{b}}$ |
| $\mathrm{EC}_{\mathrm{b}}$ | Equivalence Class broken | N | Noise |
| $\mathrm{V}_{\mathrm{EC_i}}$ | Variance for $\mathrm{EC}_{\mathrm{i}}$ | M | Total number of record in an EC |
| $\mathrm{MS}_{\mathrm{n}}$ | Max frequency of $\mathrm{A}_{\mathrm{i}}^{\mathrm{s}}$ in an $\mathrm{EC}_{\mathrm{n}}$ | $\mathrm{MS}_{\mathrm{c}}$ | Max frequency of $\mathrm{A}_{\mathrm{i}}^{\mathrm{s}}$ in an $\mathrm{EC}_{\mathrm{c}}$ |
| $\mathrm{MS}_{\mathrm{n-1}}$ | Max frequency of $\mathrm{A}_{\mathrm{i}}^{\mathrm{s}}$ in an $\mathrm{EC}_{\mathrm{n\text{-}1}}$ | $\mathrm{MS}_{\mathrm{b}}$ | Max frequency of $\mathrm{A}_{\mathrm{i}}^{\mathrm{s}}$ in an $\mathrm{EC}_{\mathrm{b}}$ |
| P | Places used in formal modeling | $\mathrm{G}_{\mathrm{i}}^{\mathrm{qi}}$ | QI-group at index *i* |
| $\varphi$ | Data Types in formal modeling | | |

**Definition 4.** *p-sensitive k-anonymity [19]: The masked microdata table T′ is p-sensitive k-anonymous if it is k-anonymous and each EC in T′ has at least p distinct $A^s$ values.*

$$\mathrm{iff}\ |\forall\{\mathrm{A}_{\mathrm{i}}^{\mathrm{qi}} \times \mathrm{t}(\mathrm{A}_{\mathrm{in}}^{\mathrm{qi}})\}| \in \mathrm{T}' \geq k \wedge$$
$$(\ \forall \mathrm{G} : \{\mathrm{A}_{\mathrm{i}}^{\mathrm{qi}} \times \mathrm{A}_{\mathrm{i}}^{\mathrm{s}}\} \in \mathrm{T}' \bullet \mathrm{A}_{\mathrm{n}}^{\mathrm{s}} \leftarrow \mathrm{Count}(\mathrm{Dist}(\mathrm{A}_{\mathrm{i}}^{\mathrm{s}})) \geq p\,)$$

*where G represents an EC that already satisfies k-anonymity and is a set of $A_i^s$ and $A_i^{qi}$. The value of $A_n^s$ must be equal to or greater than p, where $A_n^s$ represents distinct $A^s$ values in an EC.*

**Definition 5.** *Categorical similarity attack: If an adversary knows that the l-diverse modified microdata T′ (satisfying k-anonymity and l-diversity) has sensitive values belong to the single sensitive category in an EC from a p distinct $A^s$ categories.*

**Definition 6.** *Sensitive variance attack: The privacy leakage in an EC due to the low variability of sensitive values from p distinct $A^s$ categories.*

**Definition 7.** *High-Level Petri Nets (HLPN) [37]: The behavior of the system with its mathematical properties are modeled specifically via HLPN. An HLPN is a combination of 7-tuples $N = (P, T, F, \varphi, R_n, L, M_0)$, where P represented by circles are the set of places. T is the set of transitions in the system represented by rectangular boxes, such that $P \cap T = \varnothing$. F represents the flow relations such that $F \subseteq (P \times T) \cup (T \cup P)$. $\varphi$ maps places P to the data types. $R_n$ represents the rules or properties for transitions that verify the correctness of the underlying system. L represents labels on F, and $M_0$ is the initial marking.*

The following section reviews the $p^+$-sensitive *k*-anonymity model, to highlight its shortcomings concerning sensitive variance or an S-Variance attack.

## 4. Problem Statement

Definitions 8 and 9 describe the $p^+$-sensitive *k*-anonymity and $(p,\ \alpha)$-sensitive *k*-anonymity models [17], respectively.

**Definition 8.** *$p^+$-sensitive k-anonymity [17]: A masked microdata $T'$, fulfills k-anonymity and for each $A^s$ value belongs to distinct categories must be equal to or greater than p for each EC in $T'$.*

$$(\forall G : \{A_i^{qi} \times A_i^{s}\} \in G \cup C \wedge \\ \forall C \in G \bullet C_n \leftarrow \mathrm{Count}(\mathrm{Dist}(C)) \geq p)$$

*where C depicts $A^s$ values categorizations that already fulfill a p-sensitive k-anonymous approach. $C_n$ represents distinct categories in Table 3 [17] and must be equal to or greater than* p*. Table 4a obtained from Table 1a, shows $p^+$-sensitive k-anonymity model in which p = 2, k = 4 and c = 2. The ECs column in Table 4a is not part of a published table.*

**Table 3.** Category table.

| Category ID | Sensitive Values |
|---|---|
| 1 | HIV, Cancer |
| 2 | Hepatitis, Phthisis |
| 3 | Asthma, Obesity |
| 4 | Indigestion, Flu |

**Table 4. a.** $2^+$-Sensitive 4-Anonymous. **b.** (3,1)-Sensitive 4-Anonymous.

| ECs | ID | Age | Zip Code | Country | Disease |
|---|---|---|---|---|---|
| EC1 | 1 | =< 40 | 14204-14247 | America | HIV |
| | 2 | =< 40 | 14204-14247 | America | Cancer |
| | 3 | =< 40 | 14204-14247 | America | Flu |
| | 4 | =< 40 | 14204-14247 | America | Indigestion |
| EC2 | 5 | >= 40 | 13073-14066 | **** | Hepatitis |
| | 6 | >= 40 | 13073-14066 | **** | Phthisis |
| | 7 | >= 40 | 13073-14066 | **** | Asthma |
| | 8 | >= 40 | 13073-14066 | **** | Obesity |
| EC3 | 9 | =< 40 | 14203-14247 | **** | HIV |
| | 10 | =< 40 | 14203-14247 | **** | Cancer |
| | 11 | =< 40 | 14203-14247 | **** | Flu |
| | 12 | =< 40 | 14203-14247 | **** | Flu |

(**a**)

| ID | Age | Zip Code | Country | Disease |
|---|---|---|---|---|
| 1 | =< 40 | 14205-14247 | **** | HIV |
| 2 | =< 40 | 14205-14247 | **** | HIV |
| 3 | =< 40 | 14205-14247 | **** | Cancer |
| 4 | =< 40 | 14205-14247 | **** | Flu |
| 5 | >= 40 | 13073-14066 | **** | Hepatitis |
| 6 | >= 40 | 13073-14066 | **** | Phthisis |
| 7 | >= 40 | 13073-14066 | **** | Asthma |
| 8 | >= 40 | 13073-14066 | **** | Obesity |
| 9 | =< 40 | 14203-14247 | America | Cancer |
| 10 | =< 40 | 14203-14247 | America | Flu |
| 11 | =< 40 | 14203-14247 | America | Flu |
| 12 | =< 40 | 14203-14247 | America | Indigestion |

(**b**)

**Definition 9.** $(p, \alpha)$*-sensitive k-anonymity* [17]*: A modified microdata table $T'$ that fulfills the k-anonymity property and there must be p distinct sensitive attribute values in each QI-group having a minimum weight of at least $\alpha$.*

$$(\forall \mathrm{G} : \{\mathrm{A}_{\mathrm{i}}^{\mathrm{qi}} \times \mathrm{A}_{\mathrm{i}}^{\mathrm{s}}\} \in \mathrm{T}' \wedge \ \mathrm{A}_{\mathrm{n}}^{\mathrm{s}} \ \geq \mathrm{p} \ \wedge \mathrm{w_c} \ \geq \alpha)$$

*where G represents all groups in masked micro table $T'$ that already fulfill the p-sensitive k-anonymity property. Weight should be assigned to each category and each sensitive value p must have weight in each category i.e., $w_c$ that must be at least $\alpha$. Table 4b obtained from Table 1a, shows $(p, \alpha)$-sensitive k-anonymity.*

The sensitive variance and categorical similarity attacks have minor difference concerning the variability of $\mathrm{A}^{\mathrm{s}}$ in an EC. The sensitive variance attack is more powerful than categorical similarity attack, i.e., *categorical similarity attack* $\in$ *sensitive variance attack*. Therefore, the attribute disclosure through the sensitive variance attack automatically covers the disclosures through the categorical similarity attack. The EC2 and EC3 in Table 4a obtained through the $p^+$-sensitive $k$-anonymous approach have categorical similarity and sensitive variance attacks and are explained in Table 5. Table 5 shows the variance calculation for these ECs, where a high variance for more diverse EC2 and small variance for less diverse EC3 can be seen.

**Table 5.** Variance calculation for different equivalence classes (ECs) in Table 4a.

| **EC2** | | | | | | **EC3** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sensitive Values | $x$ | $f$ | $x^2$ | $f * x$ | $f * x^2$ | Sensitive Values | $x$ | $f$ | $x^2$ | $f * x$ | $f * x^2$ |
| Hepatitis | 1 | 1 | 1 | 1 | 1 | Flu | 1 | 2 | 1 | 2 | 2 |
| Phthisis | 2 | 1 | 4 | 2 | 4 | Cancer | 2 | 1 | 4 | 2 | 4 |
| Asthma | 3 | 1 | 9 | 3 | 9 | HIV | 3 | 1 | 9 | 3 | 9 |
| Obesity | 4 | 1 | 16 | 4 | 16 | | | $N = \sum f = 4$ | | $\sum fx = 7$ | $\sum fx^2 = 15$ |
| | | $N = \sum f = 4$ | | $\sum fx = 10$ | $\sum fx^2 = 30$ | | | | | | |
| Variance $(\sigma^2)$ | | $\left(\frac{\sum \mathrm{fX}^2}{\mathrm{N}} - (\frac{\sum \mathrm{fX}}{\mathrm{N}})^2\right) = \left(\frac{30}{4} - (\frac{10}{4})^2\right) = 1.25$ | | | | Variance $(\sigma^2)$ | | $\left(\frac{\sum \mathrm{fX}^2}{\mathrm{N}} - (\frac{\sum \mathrm{fX}}{\mathrm{N}})^2\right) = \left(\frac{15}{4} - (\frac{7}{4})^2\right) = 0.69$ | | | |

To calculate the variance of the ECs, an ordered weight is given to the $\mathrm{A}^{\mathrm{s}}$ values in such a way that the higher the frequency ($f$), the lower the weight ($x$) will be. For example, consider EC3 in Table 4a, i.e., Flu = 2, Cancer = 1, HIV = 1. The numeric value against each sensitive value represents its frequency occurrence in EC3. If an EC, e.g., EC2 is fully diverse i.e., size 4 and 4-diverse, then the order weight will be Hepatitis = 1, Phthisis = 1, Asthma = 1, Obesity = 1. In EC2, because of having a single occurrence for each $\mathrm{A}^{\mathrm{s}}$ value, has a higher variance than EC3.

An adversary, using the category table (Table 3), can analyse the ECs in Tables 4a and 4b published in [17]. The variability in some of the ECs is low concerning the category table. Therefore, the adversary can isolate the sensitive values that belong to a specific category and hence to individual records, and thus breaches the identity of an individual.

### *Critical Review of $p^+$-Sensitive k-Anonymity Model*

We formally modeled the $p^+$-sensitive $k$-anonymity algorithm to check its invalidation concerning a sensitive variance attack. The detail formal verification of the working of $p^+$-sensitive $k$-anonymity privacy model along with its properties is given in [18] from Rule 1 to Rule 7, which gets original data input from the end-user and processes it. The sensitive variance attack over the $p^+$-sensitive $k$-anonymity model is shown in Figure 1, where the arrow heads show the data flow. Table 6 shows variable types and their descriptions. The places $P$ and its description are shown in Table 7. The attacker model in Figure 1 consists of three entities: the end-user, the adversary, and the trusted data publisher.

**Table 6.** Types used in high-level Petri nets (HLPN) for p+-sensitive k-anonymity.

| Data Types | Description |
|---|---|
| $k$ | User input for $k$-anonymity |
| $p$ | $p$-sensitivity numeric value |
| C | Distinct categories set |
| Condition | Boolean value 1 or 0 |
| $S_n$ | Total distinct $A^s$ values |
| $C_n$ | Total distinct categories |
| $A_i^{si}$ | Sensitive Attribute for $i^{th}$ end user |
| $A_i^{id}$ | Identifier attribute for $i^{th}$ end user |

**Table 7.** Data-types, places, and their mapping.

| Places | Description |
|---|---|
| $\varphi$ **(MT)** | $\mathbb{P}\ (A^{qi} \times A^s \times A^{id})$ |
| $\varphi$ **(MMT)** | $\mathbb{P}\ (A^{qi} \times A^s \times k)$ |
| $\varphi$ **(KLevel)** | $\mathbb{P}\ (k)$ |
| $\varphi$ **(CondTF)** | $\mathbb{P}$ (Condition) |
| $\varphi$ **(Gi)** | $\mathbb{P}\ (A^{qi} \times A^s \times k)$ |
| $\varphi$ **(ds)** | $\mathbb{P}\ (A^s)$ |
| $\varphi$ **(CountDs)** | $\mathbb{P}\ (S_n)$ |
| $\varphi$ **(Gi′)** | $\mathbb{P}\ (A^{qi} \times A^s \times k \times C)$ |
| $\varphi$ **(PLevel)** | $\mathbb{P}\ (p)$ |
| $\varphi$ **(CompC)** | $\mathbb{P}\ (C_n)$ |
| $\varphi$ **(Publish Data)** | $\mathbb{P}\ (A^{qi} \times A^s)$ |
| $\varphi$ **(BK)** | $\mathbb{P}\ (A^{id} \times A^{qi})$ |
| $\varphi$ **(SA Disc)** | $\mathbb{P}\ (A_i^{qi} \times A_i^{si} \times A_i^{id})$ |

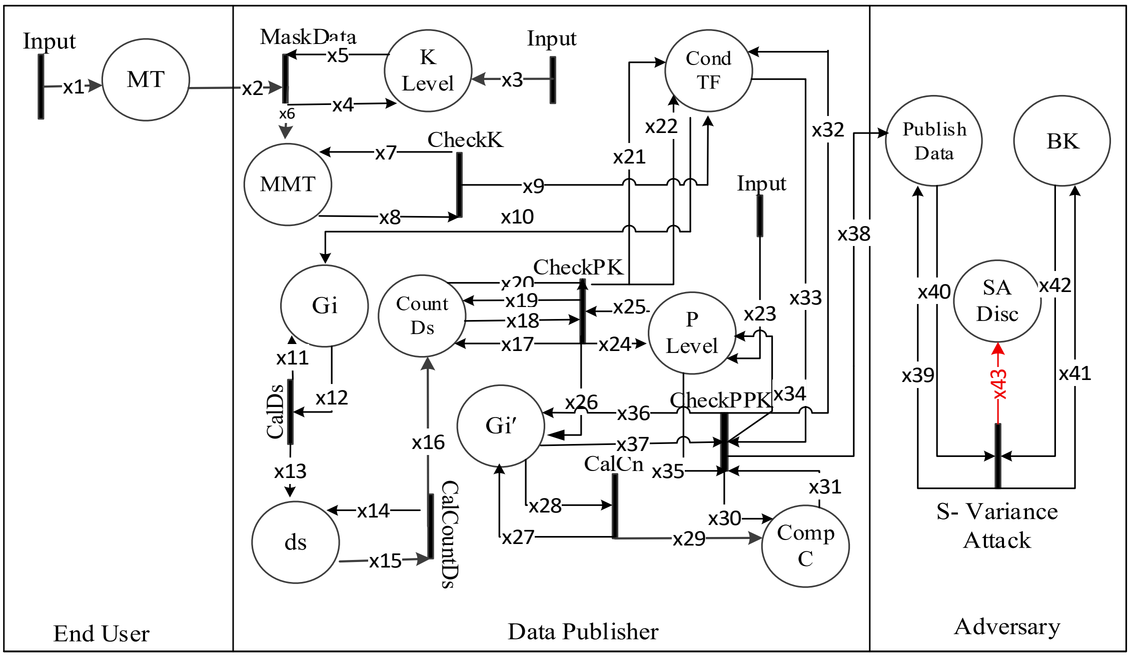

**Figure 1.** HLPN for $p^+$-sensitive $k$-anonymity attack model.

In Figure 1, Transitions $T$, which are input to the HLPN model, consist of patients' records (original data). A trusted data publisher further processes the data to minimize an attribute disclosure risk. Generalization and removing identifying attributes transform the data into masked data. After generalization, the masked microdata table is ready to be published. An adversary then exploits the published data for its benefits.

In this paper, the first seven rules in [18] are outlined briefly. For input *k*, the data publisher processes the original data to perform data generalization via the Generalize() function and each EC is stored at place the micro mask table (MMT); The publisher confirms the *k*-anonymity condition. If successful, *Condition* variable is set to *true*. For each EC, the Dist() function calculates the distinct $A^s$ values and stores its count at place ds. To further process the array of t $A^s$, the Count() function counts the $S_n$ and stores it at place Count Ds. Before the calculation of $C_n$, *p*-sensitive *k*-anonymity is verified in masked data. Transition *CheckPK* checks at least p distinct $A^s$ values in each EC in the whole table. *PLevel* stores the input transition *p* value for comparison. Apart from the checking condition for *k*-anonymity, another checking for *p* value is done. If it returns *true* it means the data already fulfills *k*-anonymity. This concludes a successful transition, ensures the *p*-sensitive *k*-anonymous property. Next, computing $A^s$ values categories using function Get_Cat(). Both $A^s$ values and categories are stored at place Gi′ for further processing. Actual improvement to the prior model and source for *p*+-sensitive *k*-anonymity is the transition *CheckPPK*. Distinct categories are calculated in a column, using the sensitive values. Comp C stores this 'number' of distinct categories. The $C_n$ involved in each EC is checked with transition *CheckPPK* to confirm that there must be at least *p* distinct categories. The minimum value for *p* is 2. The *p*+-sensitive *k*-anonymity properties are fulfilled if the *condition* variable returns *true*.

The $p^+$-sensitive *k*-anonymity model is highly vulnerable against a sensitive variance attack. The main reason is the existence of non-diverse (low variance) $A^s$ values similar to 'Flu' in Tables 4a and 4b, and 'HIV' in Table 4b. In Rule (1) through function S-Variance_ Attck(), an adversary performs an attack on the released data using some external source of information, i.e., BK. In Rule (1):

$$\begin{aligned} \boldsymbol{R}(\text{Attack}) = {} & \forall\ \text{i40}\ \in \text{x40},\ \text{i42}\ \in \text{x42},\ \text{i43} \in \text{x43}\ ,\ \forall\ \text{i2} \in \text{x2}| \\ & \text{S} - \text{Variance\_Attck}(\text{i40}[2], \text{i42}[2]) \rightarrow \text{i43}[2] = \text{i2}[1]\ \wedge\ \text{i43}[2] = \text{i2}[3] \end{aligned}$$

The adversary takes the union of the published data with the external information and BK to plot an EC. In this way, specific individuals correspond to some specific ECs that belong to homogenous categories and hence sensitive values from a specific category disclose an individual. Therefore, a sensitive variance attack occurs due to low variance in corresponding ECs.

## 5. The proposed $\theta$-Sensitive *k*-Anonymity Privacy Model

### 5.1. Threshold $\theta$-Sensitivity

The goal of the proposed $\theta$-Sensitive *k*-anonymity privacy model is to prevent the attribute disclosure of the individual records in MT, collected through the IoT [2–6] enabled devices. Each EC in MT must satisfy the threshold $\theta$ value. The $\theta$-Sensitivity, is the product of variance ($\sigma^2$) and Observation 1 ($\mu$) as shown in Equation (1).

$$\theta = \text{Variance of a fully diverse EC } (\sigma^2) * \text{Observation } 1(\mu) \tag{1}$$

The variance value represents the diversity in an EC. High variance means high diversity in an EC and vice versa, since achieving 100% diversity is almost impossible in all cases. However, the variance-based optimal frequency distribution of $A^s$ values with some fixed amount of noise addition achieves an enhanced data privacy in an EC. The proposed method in this paper is simple and effective. During examining each EC, if the variance of an EC is greater than $\theta$ i.e., fully diverse, the next EC is examined. Otherwise, the variance for the same EC is increased by swapping the $A^s$ values from the successor ECs or by adding some noise records, to make it above $\theta$. Because of the required noise addition, our proposed model implies $\varepsilon$-differential privacy [22,23] but the proposed approach is a syntactic anonymization [9] approach.

### 5.1.1. Variance ($\sigma^2$)

The variance calculation in Table 5 for ECs depicts the variability in a numerical form. To standardize the $\theta$ value for different size ECs, to prevent the sensitive variance attack, initially, we consider a fully diverse EC, e.g., if EC size = 2 variance = 0.25, if EC size = 3 variance = 0.67, if EC size = 4 variance = 1.25, if EC size = 5 variance = 2, and so on, then multiplying the variance with an observed value from Observation 1 ($\mu$).

### 5.1.2. Observation 1 ($\mu$)

A decimal multiplied part: *Observation 1 ($\mu$)*, for getting $\theta$, the threshold value has full control over the EC diversity. During the simulation in Python, different values for $\mu$ were checked to get a suitable $\theta$ value. After executing the dataset for different $k$ size ECs, the values of $\mu$ in the range of 0.5 to 0.9 were concluded. A smaller observed $\mu$ value results in the frequent repetition of sensitive values in an EC, and higher observed value produces a more diverse EC. However, "what observed value should be chosen for different size ECs?", is explained below.

Consider again, the $2^+$-Sensitive 4-anonymous Table 4a, EC2 variance = 1.25, and EC3 variance = 0.69. The difference is because of the duplicated sensitive value i.e., Flu, in EC3. We propose an efficient way of removing the frequency repetition of sensitive values to achieve a more diverse EC. For this, we calculated the $\theta$ value. For example, consider a fully diverse EC of size 4 with variance = 1.25 and multiply it with an observed value, ranges between 0.5 and 0.9. Since, 1.25 * 0.5 = 0.625 is less than 0.69 and 1.25 * 0.6 = 0.75, which is greater than 0.69. The difference between the two values i.e., 1.25 and 0.69, is because of only one duplicated value "Flu". Thus, it depends on privacy requirements and the level of diversity we are interested to achieve. In this paper, we perform a very strict $\theta$ calculation to get fully diverse ECs. Therefore, for example in the implementation part of the proposed Algorithm 1, we multiply a variance of 4 size EC with an observed value $\mu$= 0.6 to have a fully diverse EC. The same technique is applied to all other ECs as well. The $\theta$ obtained in this way in line 8 of the proposed Algorithm 1 in Section 5.2, is then checked in the conditional part at line 10 inside a loop to check all ECs concerning $\theta$ requirements.

**Definition 10.** *$\theta$-Sensitive k-anonymity: The modified microdata table $T'$ fulfills $\theta$-sensitive k-anonymity, if it fulfills k-anonymity and for each EC in $T'$, the variance for each EC must be at least $\theta$.*

$$\text{iff } |\forall\{\mathrm{A}_{\mathrm{i}}^{\mathrm{qi}} \times \mathrm{t}(\mathrm{A}_{\mathrm{in}}^{\mathrm{qi}})\}| \in \mathrm{T}' \geq k \wedge \left(\forall \mathrm{G} : \{\mathrm{A}_{\mathrm{i}}^{\mathrm{qi}} \times \mathrm{A}_{\mathrm{i}}^{\mathrm{s}}\} \in \mathrm{T}' \bullet \mathrm{A}_{\mathrm{n}}^{\mathrm{s}} \leftarrow \mathrm{Count}(\mathrm{Dist}(\mathrm{A}_{\mathrm{i}}^{\mathrm{s}})) \geq \theta\right)$$

*where G represents a QI-group or EC that already satisfies k-anonymity and is a set of $A_i^s$ and $A_i^{qi}$. The value of $A_n^s$ must be equal to or greater than p, where $A_n^s$ is the number of distinct sensitive values in a QI-group. The proposed $\theta$-sensitive k-anonymity model produces the anonymized Table 8a (with noise) from the original microdata Tables 1a and 8b (without noise) from Table 4b. The $A^{qi}$ values in Tables 8a and 8b are generalized through local recoding (bottom-up generalization) which improves the utility of the anonymized data. The 4-diverse ECs in Tables 8a and 8b have sensitive values from a minimum of three different sensitive categories in Table 3. Therefore, these tables have more attribute privacy and are more protected from a sensitive variance attack.*

**Table 8.** **a**. $\theta$-sensitive 4-anonymous (with noise). **b**. $\theta$-sensitive 4-anonymous (without noise).

| ID | Age | Zip Code | Country | Disease |
|---|---|---|---|---|
| 1 | =< 40 | 14054-14247 | America | HIV |
| 2 | =< 40 | 14054-14247 | America | Cancer |
| 3 | =< 40 | 14054-14247 | America | Hepatitis |
| 4 | =< 40 | 14054-14247 | America | Obesity |
| 5 | >= 40 | 13073-14243 | Asia | HIV |
| 6 | >= 40 | 13073-14243 | Asia | Phthisis |
| 7 | >= 40 | 13073-14243 | Asia | Asthma |
| 8 | >= 40 | 13073-14243 | Asia | Flu |
| 9 | =< 40 | 14063-14247 | America | Cancer |
| 10 | =< 40 | 14063-14247 | America | Flu |
| 11 | =<40 | 14063-14247 | America | Flu |
| 12 | =<40 | 14063-14247 | America | Indigestion |
| 13 | =<40 | 14063-14247 | America | Obesity |

(**a**)

| ID | Age | Zip Code | Country | Disease |
|---|---|---|---|---|
| 1 | =< 40 | 14054-14247 | America | Hepatitis |
| 2 | =< 40 | 14054-14247 | America | HIV |
| 3 | =< 40 | 14054-14247 | America | Cancer |
| 4 | =< 40 | 14054-14247 | America | Flu |
| 5 | >= 40 | 13073-14243 | Asia | HIV |
| 6 | >= 40 | 13073-14243 | Asia | Phthisis |
| 7 | >= 40 | 13073-14243 | Asia | Asthma |
| 8 | >= 40 | 13073-14243 | Asia | Flu |
| 9 | =< 40 | 14063-14247 | America | Cancer |
| 10 | =< 40 | 14063-14247 | America | Obesity |
| 11 | =< 40 | 14063-14247 | America | Flu |
| 12 | =< 40 | 14063-14247 | America | Indigestion |

(**b**)

### *5.2. The Proposed $\theta$-Sensitive k-Anonymity Algorithm*

The proposed $\theta$-sensitive $k$-anonymity algorithm starts execution by checking the $k$ size to create an EC (minimum cardinality $k = 2$), at line 3. The algorithm can be executed on different size of $k$. However, if the minimum cardinality fails, the condition becomes $false$ and jumps to line 50. If it is *true*, the $for$ loop works from line 5 to 7, to calculate the variance for each $m$ size $\mathrm{G}_{\mathrm{i}}^{\mathrm{qi}}$ or $\mathrm{EC}_{\mathrm{i}}$ that belongs to $k$-anonymous ECs and assigns them to an array i.e., $\mathrm{V}_{\mathrm{EC}_{\mathrm{i}}}$.

Line 8 multiplies an average observed value $\mu$ and variance $\sigma^2$ for an EC to get a threshold $\theta$ (i.e., Equation (1)). This $\theta$ value ensures the maximum level of diversity in an EC. $\theta$ mainly depends on $\mu$. If $\mu$ is smaller for an EC, low diverse EC will be obtained and vice versa. What level of diversity we want to have in an EC is completely controlled by $\mu$. Deeply observing $l$-diversity [15] and $t$-closeness [16] and performing experiments while executing the algorithm in Python, the $\mu$ value is kept to achieve maximum diversity. The algorithm starts working from lines 9–49, which checks the obtained variances against user input $k$ for each $m$ size EC to $\theta$. At line 10, if $\mathrm{V}_{\mathrm{EC}}$ is greater than $\theta$, line 46 is executed and the algorithm moves on to next EC. If it is less than $\theta$, the current EC is named as $\mathrm{EC}_{\mathrm{c}}$, and the next index EC is named as $\mathrm{EC}_{\mathrm{b}}$. Lines 12–45, each part inside *if* statement has two major functionalities; swapping and require noise addition.

The *else* part of an *if* statement executes the ECs from first till $\mathrm{EC}_{\mathrm{n-1}}$, and its first part processes the last EC (lines 13–25). At line 12, if $\mathrm{EC}_{\mathrm{c}}$ is the last class, i.e., $\mathrm{EC}_{\mathrm{n}}$, then from $\mathrm{A}^{\mathrm{s}}_{\mathrm{EC}_{\mathrm{n}}}$ the value of $\mathrm{MS}_{\mathrm{n}}$ is calculated. Similarly, the value of $\mathrm{MS}_{\mathrm{n-1}}$ is calculated from $\mathrm{EC}_{\mathrm{n-1}}$. At line 15, a crossCheck() function checks the existence of most frequent $\mathrm{A}^{\mathrm{s}}$ that does not exist in each other ECs. The swap() function may be executed. The purpose of the cross-check is not to further increase or decrease $\mathrm{V}_{\mathrm{EC}_{\mathrm{n-1}}}$ because it has already been processed by the *else* part of the current *if* statement. This function is for the last EC to increase its diversity. If any of the $\mathrm{A}^{\mathrm{s}}$ value from $\mathrm{EC}_{\mathrm{n}}$ exists in $\mathrm{EC}_{\mathrm{n-1}}$ or vice versa, the swapping at line 17 will not be performed. If swapping is performed, $\mathrm{V}_{\mathrm{EC}_{\mathrm{n}}}$ is calculated to check with $\theta$ (line 20).

If $V_{EC_n}$ is still less than $\theta$, then the algorithm jumps to line 43 to add a distinct $A^s$ value as noise to increase its variance and to achieve a high diversity.

To process the first EC until the $EC_{n-1}$, the else part of *if* statement executes (line 12). The algorithm finds an $EC_b$ with $\theta$ greater than $EC_c$ (lines 27–31). The *if* statement checks $EC_b$ *found* condition, when it is satisfied, then a function mfsv() is executed on both $EC_c$ and $EC_b$, which calculates the most frequent sensitive values in both ECs. Before swapping the values for $MS_c$ and $MS_b$, a function backCheck() checks the existence of $MS_b$ in $EC_c$, which is an EC ahead of $EC_b$. If the value of $MS_b$ exists in $EC_c$, then that MS value is removed from a temporary array in Algorithm 1.

**Algorithm 1: $\theta$-sensitive $k$-anonymity**

```
Input: Microdata Table (MT)
Output: : $\theta$ − sensitive $k$ − anonymous table (MMT)

1   Procedure : $\theta$ − sensitive $k$ − anonymity (MMT, $\theta$, k)
2   Let $k \subseteq$ MMT
3   if|k| $\geq 2$ then
4       Condition = true;
5           for each m size EC in $G_i^{qi}$ : $\{A_i^{qi} \times A_i^s\} \in k$ do        ▸ $G_i^{qi}$ set, consists of $A_i^{qi}$&$A_i^s$
6               $V_{EC_i} \leftarrow$ Compute vari($A^s_{EC_i}$)                      ▸ vari($A^s_{EC_i}$), calculate variance for each m size EC.
7           end for
8           $\theta \leftarrow \mu * \sigma^2$                                     ▸ $\theta$, required threshold
9           for each m size $EC_i$ in $G_i^{qi}$ : $\{A_i^{qi} \times A_i^s\} \in k$ do      ▸ $G_i^{qi}$ set, consists of $A_i^{qi}$ and $A_i^s$
10              if $V_{EC_c} < \theta$ then
11                  $EC_b \leftarrow EC_c + 1$
12                  if $EC_n = EC_c$
13                      $MS_n \leftarrow$ Compute mfsv($A^s_{EC_n}$)                 ▸ mfsv(), max frequent $A^s_{EC_n}$
14                      $MS_{n-1} \leftarrow$ Compute mfsv($A^s_{EC_{n-1}}$)         ▸ mfsv(),max frequent $A^s_{EC_{n-1}}$
15                      notExist $\leftarrow$ crossCheck($MS_{EC_n}$,$MS_{EC_{n-1}}$)  ▸ crossCheck(), check both side existence
16                      if notExist
17                          swap($MS_n$, $MS_{n-1}$)                             ▸ swap(), last and 2nd last ECs MS values
18                      end if
19                      $V_{EC_n} \leftarrow$ Compute vari($A^s_{EC_n}$)
20                      if $V_{EC_n} < \theta$
21                          Break
22                          jump to else part of condition line 43
23                      else
24                          Break
25                      end if
26                  else
27                      for $EC_b$till $EC_n$in $G_i^{qi}$ : $\{A_i^{qi} \times A_i^s\} \in K$
28                          if $V_{EC_b} > \theta$
29                              Break loop
30                          end if
31                          Break loop
32                          if $EC_b$ = found
33                              $MS_c \leftarrow$ Compute mfsv($A^s_{EC_c}$)         ▸ mfsv(), max frequency $A^s_{EC_c}$
34                              $MS_b \leftarrow$ Compute mfsv($A^s_{EC_b}$)         ▸ mfsv(), max frequency $A^s_{EC_b}$
35                              $MS_b \leftarrow$ backCheck($MS_{EC_c}$, $MS_{EC_b}$)  ▸ backCheck() find MS value in $MS_{EC_b}$, not exists in $MS_{EC_c}$
36
37                              swap($MS_c$, $MS_b$)                             ▸ swap(), exchange MS values
38                              $V_{EC_c} \leftarrow$ Compute vari($A^s_{EC_c}$)     ▸ vari(), again compute variance
39                              if $V_{EC_c} > \theta$
40                                  $EC_c + = 1$
41                              end if
42                          else
43                              NS $\leftarrow$ Compute addNoise($A^s_{EC_c}$)       ▸ addNoise(), until variance>$\theta$
44                          end if
45                      end if
46                  else
47                      $EC_c + = 1$
48                  end if
49              end for
50  else
51      Condition = false;
52  end if
```

$MS_{EC_b}$ and next MS in same $EC_b$ is checked with $MS_c$. This process continues until it finds a $A^s$ value in $MS_{EC_b}$ that do not exist in $MS_c$. Line 37 then swaps these two MS values along with their corresponding records. Two important purposes are achieved through this swap function. First, reducing the frequency of repeated $A^s$ and second, increasing diversity in $EC_c$ which results in increasing $V_{EC_c}$. The $V_{EC_c}$ is again calculated and is checked with $\theta$, if it is greater than $\theta$, counter for $EC_c$ moves to the next EC.

Here, the absence of the *else* statement adds noise instantly, in a situation when the variance is less than $\theta$, because more than one swapping for a specific $EC_c$ is possible. We add noise only once after completely checking the frequency of each $A^s$ in an EC. For example, if to produce a 4-anonymous EC table from Table 1a, after one swapping e.g., 'HIV' swaps with 'Obesity', the resulting EC1 in Table 8a will become 3-diverse and its variance will not meet $\theta$, the *else* part might add noise to increase variance even though there is a duplicated $A^s$ 'Cancer' value that still exists in $EC_c$. To reduce the frequency of the next duplicated $A^s$ value i.e., 'Cancer', by swapping it with another $A^s$ in $EC_b$ if one exists, noise is not added at this moment. This is achieved by going control back to line 10, and since this increased variance is still less than $\theta$, the procedure repeats and from an $EC_b$ a new $A^s$ is swapped with the next duplicated $A^s$ value. In this way, two swapping procedures are performed and 2-diverse $EC_c$ will become 4-diverse without adding any unnecessary noise, which results in increasing data utility and a more diverse EC.

$EC_b$ is found because of a variance greater than $\theta$, there are chances that no EC exists in a given dataset having a higher variance than $\theta$, in this case, the loop will not break (line 29). In that case, the algorithm will jump to line 43. It will add a dummy record with distinct $A^s$ value(s) via function addNoise(). Such an addition is considered as noise to the real data just like the addition of noise in differential privacy [22,23]. This algorithm performs very intelligent swapping and adds noise intelligently. The purpose of these two functions (i.e., swap() and addNoise()), is to increase the diversity keeping the utility as high as possible, which is easily achieved in our algorithm as shown in the experimental evaluation, Section 6.

The sanitized Table 4a from $p^+$-sensitive $k$-anonymity is prone to homogeneity, categorical similarity, and sensitive variance attacks, and Table 8a from $\theta$-sensitive $k$-anonymity secures the data from such attacks because of more diversity, even at the category level, i.e., the maximum value for category c is 4 through $\theta$-sensitive $k$-anonymity, where, for Table 4a, the maximum value for c is 2. Table 8a provides more protection against the categorical similarity attack. Further swapping of values is not possible in the last EC; thus, a single tuple is added as noise to increase the diversity and to prevent categorical similarity attack and sensitive variance attack. Such a small amount of noise does not highly affect the utility of the data. Table 4b is a base table to obtain Table 8b using the $\theta$-sensitive $k$-anonymity approach. Table 8b is also highly diverse at the categorical level and there are no repeated sensitive values. Thus, there is no need to add noise and to have a high value of variance. The anonymized data, both in Tables 8a and 8b, obtained through the proposed $\theta$-sensitive $k$-anonymity algorithm, have no attribute disclosure risk and are defensive against homogeneity [11], categorical similarity, and sensitive variance attacks, and even secure from skewness attacks [12].

### *5.3. Analysis of θ-Sensitive k-Anonymity Model Using Formal Modeling and Analysis*

The proposed $\theta$-sensitive $k$-anonymity model mitigates the vulnerability discussed in Section 4. Modeling the $\theta$-sensitive $k$-anonymity via HLPN has the same end-user, data publisher, and unknown adversary, as shown in Figure 2. Tables 9 and 10, respectively, show variable types and places, and their corresponding descriptions.

The $\theta$-sensitive $k$-anonymity algorithm was modeled through the HLPN rules for the microdata input. The data publisher initially verifies the $k$-value input. The original data is $k$-anonymized (bottom-up generalization) after finalizing the individual records in an EC obtained through variance calculations. In Rule (2), the $k$-anonymity masks the data. In Rule (2):

$$\boldsymbol{R}(\text{MaskData}) := \forall\ \text{i2} \in \text{x2},\ \text{i3} \in \text{x3}, \text{i4} \in \text{x4}\ |$$

$$\text{i4}[1] := \text{Mask}\ \{\text{i2}[2]\} \wedge \text{i4}[2] := \text{Mask}\{\text{i2}[3]\}\ \wedge\ \text{x4}' := \text{x4}\ \cup \{\text{i4}[1], \text{i4}[2], \text{i3}\}$$

**Table 9.** Types used in HLPN for $\theta$-sensitive $k$-anonymity.

| Data Types | Descriptions |
|---|---|
| $M$ | Size of an EC |
| Condition | Boolean value 1 or 0 |
| $\sigma$ | A float type value to define Sigma |
| $\mu$ | A float type value to define Mu |
| $\theta$ | A float type value to define Theta |
| Found $\text{EC}_\text{b}$ | Equivalence class b when it is found |
| $\text{AdjEC}_\text{c}$ | Adjust Equivalence class c |
| $\text{AdjEC}_\text{n}$ | Adjust Equivalence class n |
| $\text{VarEC}_\text{s}$ | Variance of different Equivalence classes |
| $\text{VarAdjEC}_\text{n}$ | Adjust variance for Equivalence class n |
| $\text{VarAdjEC}_\text{c}$ | Adjust variance for Equivalence class c |

**Table 10.** Mapping of data types in $\theta$-sensitive $k$-anonymity model.

| Places | Descriptions |
|---|---|
| $\varphi(\text{MT})$ | $\mathbb{P}\ (\text{A}^{\text{id}} \times \text{A}^{\text{qi}} \times \text{A}^{\text{s}})$ |
| $\varphi(\text{MMT})$ | $\mathbb{P}\ (\text{EC}_\text{c} \times \text{EC}_\text{b} \times \text{EC}_\text{n} \times k)$ |
| $\varphi(\text{KValue})$ | $\mathbb{P}\ (k)$ |
| $\varphi(\text{CondTF})$ | $\mathbb{P}\ (\text{Condition})$ |
| $\varphi(\text{Sigma})$ | $\mathbb{P}\ (\sigma)$ |
| $\varphi(\text{Mu})$ | $\mathbb{P}\ (\mu)$ |
| $\varphi(\text{Theta})$ | $\mathbb{P}\ (\theta)$ |
| $\varphi(\text{Found EC}_\text{b})$ | $\mathbb{P}\ (\text{EC}_\text{b})$ |
| $\varphi(\text{VarEC}_\text{s})$ | $\mathbb{P}\ (\text{V}_{\text{EC}_\text{c}} \times \text{V}_{\text{EC}_\text{b}} \times \text{V}_{\text{EC}_\text{n}})$ |
| $\varphi(\text{AdjEC}_\text{c})$ | $\mathbb{P}\ (\text{EC}_\text{c})$ |
| $\varphi(\text{AdjEC}_\text{n})$ | $\mathbb{P}\ (\text{EC}_\text{n})$ |
| $\varphi(\text{StrictEC}_{\text{n}-1})$ | $\mathbb{P}\ (\text{EC}_{\text{n}-1})$ |
| $\varphi(\text{VarAdjEC}_\text{n})$ | $\mathbb{P}\ (\text{V}_{\text{EC}_\text{n}})$ |
| $\varphi(\text{VarAdjEC}_\text{c})$ | $\mathbb{P}\ (\text{V}_{\text{EC}_\text{c}})$ |
| $\varphi(\text{Need Noise})$ | $\mathbb{P}\ (\text{V}_{\text{EC}_\text{c}} \times \text{A}^{\text{id}} \times \text{A}^{\text{qi}} \times \text{A}^{\text{s}})$ |
| $\varphi(\text{PublshdData})$ | $\mathbb{P}\ (\text{A}^{\text{qi}} \times \text{A}^{\text{s}})$ |
| $\varphi(\text{BK})$ | $\mathbb{P}\ (\text{A}^{\text{id}} \times \text{A}^{\text{qi}})$ |
| $\varphi(\text{SA Disc})$ | $\mathbb{P}\ (\text{A}_\text{i}^{\text{qi}} \times \text{A}_\text{i}^{\text{si}} \times \text{A}_\text{i}^{\text{id}})$ |

If an input $k$ is less than the minimum size of an EC (i.e., <2) the condition fails. For cardinality having a minimum value of 2 or above, the algorithm executes. The $k$-anonymity for *true* or *false* are depicted in Rule (3).

$$\boldsymbol{R}(\text{Check } k) := \forall\ \text{i5} \in \text{x5},\ \text{i6}\ \in\ \text{x6}\ |$$
$$\text{Count}(\text{i5}[1]) \geq \text{i5}[3] \rightarrow \text{i6}[1] := \text{TRUE}\ \vee\ \text{Count}(\text{i5}[1]) \ngeq \text{i5}[3] \rightarrow \text{i6}[1] := \text{FALSE} \wedge \text{x6}' := \text{x6} \cup \{\text{i6}[1]\}$$

The threshold $\theta$ is calculated in Rule (4). Variance for a fully diverse ECs for a specific $k$ is calculated using the var() function. The important contributed functions are swap() and addNoise() functions, through which the algorithm processes all ECs. Transition *Adjust Var* performs all these swapping and noise additions in corresponding ECs. In Rule (5), *Compute Var* transitions for the initial ECs. For the rest of the ECs, the same transition can be used in the same manner. In Rule (4):

$$\boldsymbol{R}(\text{Calc Theta}) := \forall\ \text{i10} \in \text{x10},\ \text{i11} \in \text{x11}, \text{i12} \in\ \text{x12}|$$
$$\text{i12} := \{\text{i11} * (\text{i10})^2\}\ \wedge\ \text{x12}' := \text{x12} \cup \{\text{i12}\}$$

In Rule (5):

$$\begin{aligned}\boldsymbol{R}(\text{Compute Var}) := \forall\ i8\ \in x8,\ i9 \in x9|\\ i9[1] := \text{Compute Var}(i8[1]) \wedge i9[2] := i8[2])\ \wedge x9' := x9\ \cup \{i9[1], i9[2]\}\end{aligned}$$

The $\theta$-sensitive $k$-anonymity model's main functionalities are described in Rule (6) and Rule (7). Variance in each $k$-anonymous EC with respect to $\theta$ is checked in Rule (6). If variance of $\mathrm{EC_c}$ is greater than $\theta$ (i.e., (i14[1] > i13)), move to next $\mathrm{EC_c}$ and update the value in place MMT. If the variance of $\mathrm{EC_c}$ is less than $\theta$ (i.e., (i14[1] < i13)), then transaction stops. We try to find $\mathrm{EC_b}$, and swap required available $\mathrm{A^s}$ values from $\mathrm{EC_b}$. After performing all needed swapping, if the variance of $\mathrm{AdjEC_c}$ is still less than $\theta$ (i.e., (i32 < i13)), the noise is added to increase its diversity. In Rule (6):

$$\begin{aligned}\boldsymbol{R}(\text{Check Variance}) := \forall\ i13 \in x13,\ i14 \in x14,\ i15 \in x15, i19 \in x19,\ i23 \in x23, i24 \in x24,\ i32 \in x32,\ i33 \in x33|\\ \wedge\{((\ i14[1] > i13) \rightarrow i16[1] := i15[1] + 1\\ \wedge x16' := x16[2] \cup \{i16\}) \vee ((i14[1] < i13) \rightarrow i16[2] := i15[1] + 1\\ \wedge x16' := x16[2] \cup \{i16\}) \vee ((i14[2] > i13)) \rightarrow i19 \wedge x19' := x19 \cup \{i19\}) \vee ((i24 < i13) \rightarrow i25\\ \wedge x25 := x25 \cup \{i25\}) \vee ((i32 < i13) \rightarrow i33 \wedge x33 := x33 \cup \{i33\})\}\end{aligned}$$

The proposed $\theta$-sensitive $k$-anonymity algorithm starts by processing each $k$ size EC. The function Comp mfsv( ) computes the max frequency of $\mathrm{A^s_{EC_c}}$ and $\mathrm{A^s_{EC_b}}$, named as $\mathrm{MS_{EC_c}}$ and $\mathrm{MS_{EC_b}}$, respectively. A one-way checking function: backCheck(), checks for the existence of $\mathrm{MS_{EC_b}}$ at $\mathrm{FoundEC_b}$ that do not exist in earlier $\mathrm{EC_c}$.

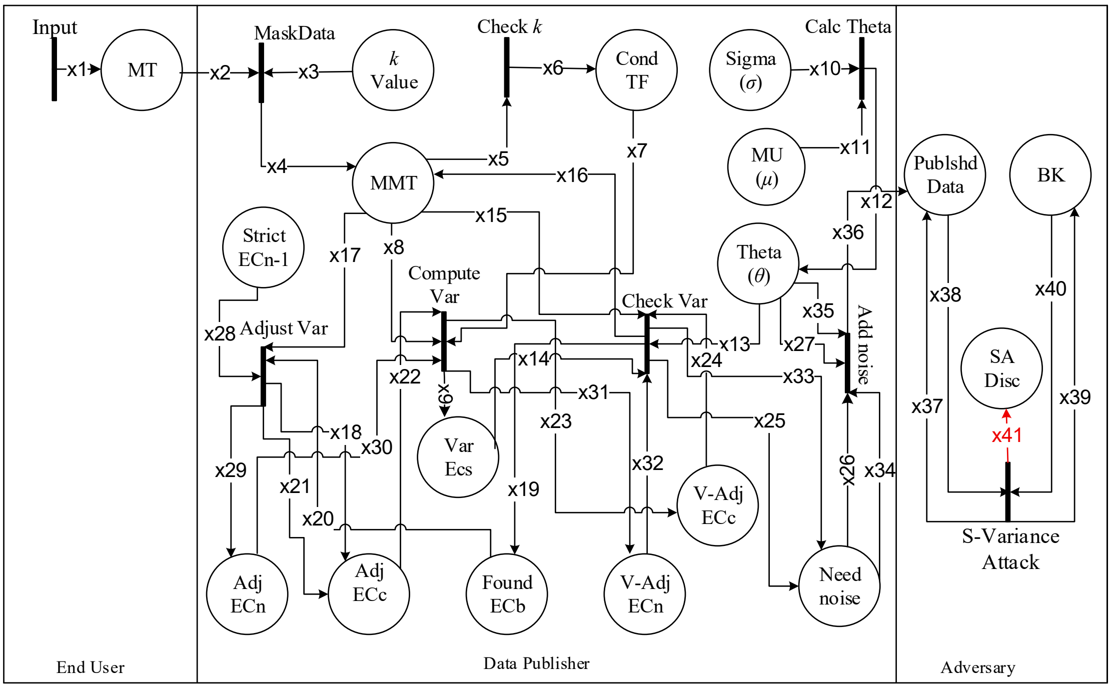

**Figure 2.** HLPN for $\theta$-sensitive $k$-anonymity.

$\mathrm{MS_b}$ is swapped with $\mathrm{MS_c}$ after the checking succeeds and is saved in place $\mathrm{AdjEC_c}$. $\mathrm{EC_c}$ minimizes the frequency of the $\mathrm{A^s}$ value and increases diversity. While processing the last EC, i.e., $\mathrm{EC_n}$, swapping is not possible in the forward direction. Thus swapping with previous EC is performed with a condition that the variance of already processed $\mathrm{EC_{n-1}}$ should not be decreased with $\theta$. The crossCheck() function confirms two-way checking, that the values for both $\mathrm{MS_n}$ and $\mathrm{MS_{n-1}}$ are distinct and it should not change the variance of $\mathrm{EC_{n-1}}$ at place $\mathrm{StrictEC_{n-1}}$ to an undesired value again. In that case, we call it strict $\mathrm{EC_{n-1}}$. In other words, in addition to increasing the diversity in $\mathrm{EC_n}$, it is also not increasing the frequency of $\mathrm{A^s}$ value at place $\mathrm{EC_{n-1}}$. Values are then swapped and are saved at place $\mathrm{AdjEC_n}$. Rule (7) shows the whole process. In Rule (7):

$$\begin{aligned}\boldsymbol{R}(\text{Adjst Var}) :=\ & \forall\ \text{i17} \in \text{x17}, \text{i20} \in \text{x20}, \text{i21} \in \text{x21}, \text{i28} \in \text{x28}, \text{i29} \in \text{x29} | \\ & (\text{i17}[1] \neq \text{i17}[3]) \rightarrow \text{Comp mfsv}(\text{i20}[1], \text{i17}[1])\ \wedge \text{True} \\ & \qquad\qquad := \text{backCheck}(\text{i20}[1], \text{i17}[1]) \\ & \wedge \text{i21} := \text{swap}(\text{i20}[1], \text{i17}[1])\ \wedge \text{x21}' := \text{x21} \cup \{\text{i21}\} \\ & (\text{i17}[1] = \text{i17}[3]) \rightarrow \text{Comp mfsv}(\text{i17}[3], \text{i28}[1])\ \wedge \text{True} \\ & \qquad\qquad := \text{crossCheck}(\text{i17}[3], \text{i28}[1]) \\ & \wedge \text{i29} := \text{swap}(\text{i17}[3], \text{i28}[1])\ \wedge \text{x29}' := \text{x29} \cup \{\text{i29}[1]\}\end{aligned}$$

If the variance of $\mathrm{AdjEC_c}$ is still less than $\theta$ (i.e., (i34[1] < i35)), a dummy record called noise is added whenever needed throughout the variance adjustment process. In Rule (8), we have given the final noise addition case for last $\mathrm{AdjEC_n}$. Its purpose is to increase the variance at a level greater than $\theta$. It will produce a highly diverse EC even if there are not enough diverse records in MMT. In Rule (8):

$$\begin{aligned}\boldsymbol{R}(\text{Add Noise}) :=\ & \forall\ \text{i34} \in \text{x34}, \text{i35} \in \text{x35}, \text{i36} \in \text{x36} | \\ & (\text{i34}[1] < \text{i35}) \rightarrow \text{i36} := \text{addNoise}(\text{i34}[2], \text{i34}[3], \text{i34}[4]) \\ & \wedge \text{x36}' := \text{x36} \cup \{\text{i36}[1], \text{i36}[2], \text{i36}[3]\}\end{aligned}$$

In Rule (9), an adversary attacks against the individual's $\mathrm{A^s}$ values. Adversary combines the already available BK (i.e., i40[2]) with the published data (i.e., i38[2]) and performs attack to disclose the patient's identity (i.e., i2[2]) and the sensitive values (i.e., i2[3]). $\theta$-sensitive $k$-anonymity model can provide better privacy protection to prevent from attribute disclosure attacks because it considers the high value of variance due to swapping and noise addition in corresponding ECs. The diversity of sensitive attribute values in ECs prevents the adversarial BK and is more effective as compared to the $p^+$-sensitive $k$-anonymity model. Therefore, the adversary did not get private information for the target individual and the attack results in a null value. In Rule (9):

$$\begin{aligned}\boldsymbol{R}(\text{S} - \text{Variance Attack}) :=\ & \forall\ \text{i38} \in \text{x38},\ \text{i40} \in \text{x40},\ \text{i41} \in \text{x41} | \\ & \text{Att\_Dis}(\text{i38}[2], \text{i40}[2]) \neq (\text{i2}[1] \cup \text{i2}[2] \cup \text{i2}[3]) \\ & (\text{i41}[2]\ \cup \text{i41}[3]) = \varnothing\end{aligned}$$

## 6. Experimental Evaluation

In this section, the experiments that were performed to show the effectiveness of the proposed $\theta$-sensitive $k$-anonymity privacy model in comparison to the $p^+$-sensitive $k$-anonymity model are described. The proposed algorithm wisely diversified the $\mathrm{A^S}$ values in a balanced way inside each EC without using the categorical approach. The utility and quality of the anonymized released data were checked with numerous quality measures.

### *6.1. Experimental Setup*

All experiments were performed on a machine with an Intel Core i5 2.39 GHz processor with 4 GB RAM, using the Windows 10 operating system. The algorithm was written in Python 3.7. We used the Adults database, which contained age, zip code, salary, and occupation attributes, which is openly accessible at the UC Irvine Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets. We considered the age, zip code, salary as $\mathrm{A^{qi}}$s and occupation as $\mathrm{A^s}$.

Experimental results show the usefulness of the proposed $\theta$-sensitive $k$-anonymity privacy model and protection against the categorical similarity attack and sensitive variance attack as compared to the $p^+$-sensitive $k$-anonymity model. The quality of the sanitized publicly released data was evaluated with four utility metrics: discernibility penalty (DCP) [18,38,39], normalized average

QI-group ($C_{AVG}$) [17,18,38], noise calculation, and query accuracy [18,33]. The execution time of both algorithms was analyzed at the end of the experiments.

### *6.2. Discernibility Penalty (DCP)*

The DCP proposed in [38] and used in [18,39] is an assignment of penalty (cost) to each tuple in the generalized data set. Through this penalty, the sanitized tuple cannot be distinguished among other tuples in the result set. Minimizing the discernibility cost is an optimal objective. The penalty for a tuple $t$ that belongs to an EC of size |EC|, i.e., t $\epsilon$ EC, will be |EC| and the penalty for each EC is $|\mathrm{EC}|^2$. The complete DCP penalty for the overall sanitized released dataset $R^*$ can be seen in Equation (2).

$$\mathrm{DCP}\ (\mathrm{R}^*) = \sum_{\mathrm{i}=1}^{|\{\mathrm{EC}\}|} |\mathrm{EC}_{\mathrm{i}}|^2 \tag{2}$$

where {EC} are the total number of ECs in $\mathrm{R}^*$. A baseline can be obtained from the most optimal DCP score calculations as shown in [10]. For example, if $k$ = 2 and the number of anonymized tuples are 10, the DCP optimal score will be $2^2 + 2^2 + 2^2 + 2^2 + 2^2 = 20$. This optimal score is called the baseline. The approach to generate groups followed in this paper was based on $k$ size, inclusive of the noise tuple(s). Higher $k$ means bigger group size, so the baseline moves up because of a high DCP score. The $p^+$-sensitive $k$-anonymity model generated groups based on $p$. It means the number of tuples can be greater than $p$ in a $k$-anonymous class. Figure 3 shows the DCS score for $\theta$-sensitive $k$-anonymity, including a comparison with $p^+$-sensitive and baseline. In comparison to $p^+$-sensitivity, the DCP score, through the proposed $\theta$-sensitive $k$-anonymity algorithm, is almost equal to the baseline, which implies that the proposed model assigned an optimal penalty to each EC and produced an optimal DCP score. The magnified subplots in Figure 3 with $k$ = 12 and $k$ = 16 for $\theta$-sensitive $k$-anonymity shows the very minor difference with baseline. This minor difference can also be seen in Table 11, with an average DCP score of 47.2 or 0.002679% with a baseline obtained from the simulation while calculating the DCP for the anonymized dataset $\mathrm{R}^*$.

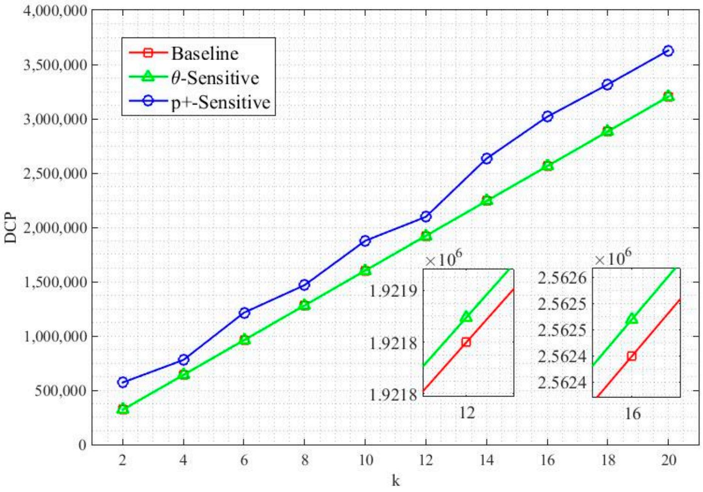

**Figure 3.** Discernibility penalty (DSP) score.

**Table 11.** DCP experiment values for each *k*.

| *k* | Baseline | $\theta$-Sensitive | $p^+$-Sensitive |
|---|---|---|---|
| 2 | 320300 | 320303 | 571227 |
| 4 | 640600 | 640605 | 778626 |
| 6 | 960900 | 960912 | 1214207 |
| 8 | 1281200 | 1281215 | 1467959 |
| 10 | 1601500 | 1601520 | 1876310 |
| 12 | 1921800 | 1921824 | 2096543 |
| 14 | 2242100 | 2242145 | 2632775 |
| 16 | 2562400 | 2562470 | 3017773 |
| 18 | 2882700 | 2882812 | 3315591 |
| 20 | 3203000 | 3203166 | 3628936 |
| Average Val. | 1761650 | 1761697.2 | 2059994.7 |
| Diff. of $\theta$ and $p^+$ avg. values with base avg. value | – | 47.2 | 298344.7 |
| Percent Closer to baseline | – | 0.002679235 | 14.65 |
| % diff. between $\theta$ and $p^+$ | – | 14.64 | – |
| This means that our proposed approach $\theta$-sensitive, k-anonymity is 14.64% better than $p^+$-sensitive k-anonymity and 0.002679% closer to the baseline. | | | |

### 6.3. Normalized Average ($C_{AVG}$)

$C_{AVG}$ is another mathematically sound measurement that measures the quality of the sanitized data by the EC average size. It was proposed in [38] and applied in [17,18]. Below in Equation (3), $C_{AVG}$ can be calculated as

$$C_{AVG} = \left(\frac{|R^*|}{|\{EC\}|}\right) \div k \quad (3)$$

where $|R^*|$ is the overall sanitized released dataset and $|\{EC\}|$ are the total number of ECs in $R^*$. Data utility and $C_{AVG}$ are inversely proportional. Low $C_{AVG}$ value indicates high information utility. The optimal goal is to have a minimum size of ECs in $R^*$. Figure 4 shows $C_{AVG}$ for $p^+$-sensitive *k*-anonymity and $\theta$-sensitive *k*-anonymity over *k*-anonymity. $p^+$-sensitive has lower data utility over small *k*, where there is a high data utility for large *k*. The proposed technique has a very balanced and sustainable utility for each input value of *k*. Thus, the proposed $\theta$-sensitive *k*-anonymity model performs efficiently for all sizes of *k*, compared to the $p^+$-sensitive *k*-anonymity model.

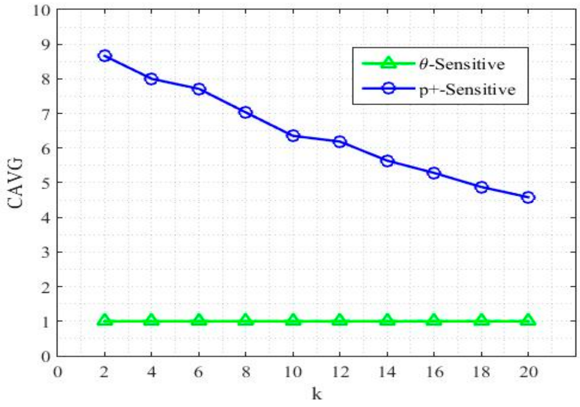

**Figure 4.** The ratio of $C_{AVG}$.

### 6.4. Noise Addition

Among different masking methods, one popular approach is the perturbation of data, i.e., noise addition. These are dummy tuples, added to the original data that helps in achieving the required diversity similar to the differential privacy [22,23]. The reason is if there are not enough $A^s$ values to swap with, especially in the second last and last ECs, the gap is filled with the noise tuples to prevent with disclosure risk. So, one of the reasons for such a good performance of the proposed model is the cost of noise addition. Figure 5 shows the number of tuples added as a noise for different values of $k$. These tuples are added to achieve the required value of the threshold $\theta$. For different values of $k$, the algorithm responds differently but the maximum number of noise tuples added for a specific value of $k$ is only six tuples. In the processed "Adult" dataset, the total number of tuples was 160,150 and only 34 noise tuples, i.e., 0.021% of the total size, were added in total. Such an amount of utility loss is negligible. This small amount of noise addition is sometimes due to get a round number when dividing the dataset size by the $k$ size input, for example, 160150/4 = 40037.5 and 160152/4 = 40038.

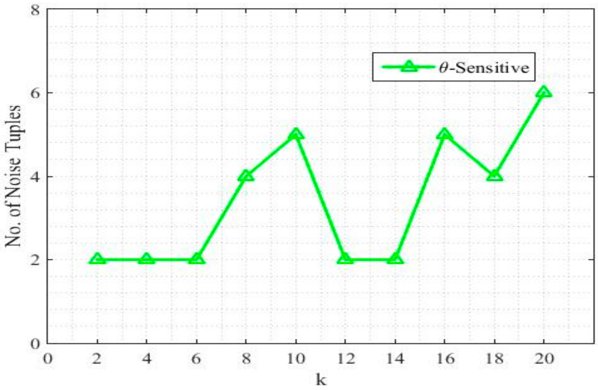

**Figure 5.** The number of noise tuples added against each $k$.

### 6.5. Query Accuracy

Query accuracy measures precision for aggregate queries to check the utility of the anonymized data. It has been used by various research works [18,33]. To answer the aggregate queries, the built-in COUNT operator is used, where $A^{qi}$s are the query predicates. Consider $R^*$ to be a sanitized release from original microdata R having maximum $m$ as $A^{qi}$s; $A_i^{qi}$ $(1 \le i \le m)$, where $D(A_i^{qi})$ is the domain of i$^{th}$ QI. The SQLQuery in Equation (4) for the COUNT query will work as

$$\text{SQLQuery} = \text{select COUNT}(*) \text{ from } R^* \text{ where } A_1^{qi} \in D(A_1^{qi}) \text{ AND } \ldots \text{ AND } A_m^{qi} \in D(A_m^{qi}) \quad (4)$$

Against each query, at least one or a few number of tuples should be selected from each EC based on query predicates. Two important parameters for query predicates are (1) query dimensionality q, and (2) the query selectivity $\vartheta$. Query dimensionality comprises of the number of QIs in query predicate while query selectivity is the number of values for each attribute $A_i$, $(1 \le i \le n)$. The query selectivity is calculated as, $\vartheta = \frac{|T_Q|}{|R|}$, where $|T_Q|$ are the output number of tuples after using query Q on relation R, and |R| are the total number of tuples in the whole dataset. Query error i.e., Error(Q), is calculated in Equation (5).

$$\text{Error}(Q) = \frac{|\text{count}(R^*) - \text{count}(R)|}{\text{count}(R)} \quad (5)$$

where count$(R^*)$ depicts result set from the COUNT query on an anonymized dataset while count(R) is the result set from the COUNT query on the original microdata. More selective queries have a high error rate.

Figure 6a shows the query error for the input value of $k$. We compare the $p^+$-sensitive $k$-anonymity and $\theta$-sensitive $k$-anonymity using the query error rate for 1000 randomly generated aggregate queries. The error rate increases for the high value of $k$ because of the high range in $A^{qi}$s. This selects a greater number of tuples than the original microdata and hence high error rate. In Figure 6b, it is depicted that the more we select tuples based on predicates, the higher the error rate will be in the anonymized data.

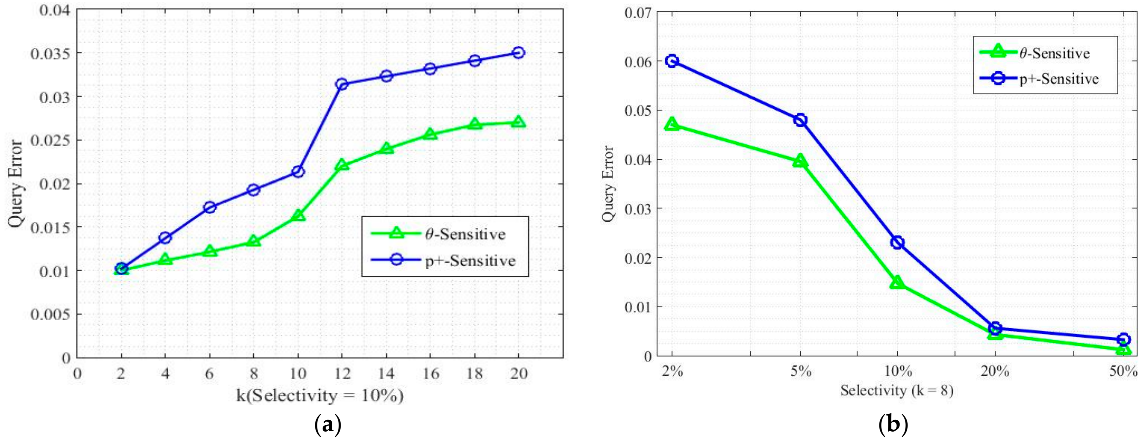

**Figure 6.** (**a**) Query error for $k$. (**b**) Query error for selectivity.

### 6.6. Execution Time

Figure 7 shows the execution time for both $p^+$-sensitive $k$-anonymity model and for the proposed $\theta$-sensitive $k$-anonymity model. The execution time for both of the algorithms increased with an increase in value of $k$ because of the increase in $A^{qi}$s generalization range. Since we did not consider the sensitive values categorization, our approach took a small amount of time to execute as compared to its counterpart. In the $\theta$-sensitive $k$-anonymity model, a higher execution time for $k = 10$, $k = 16$ and $k = 20$ was because of the time taken to add more noise tuples to achieve the required diversity.

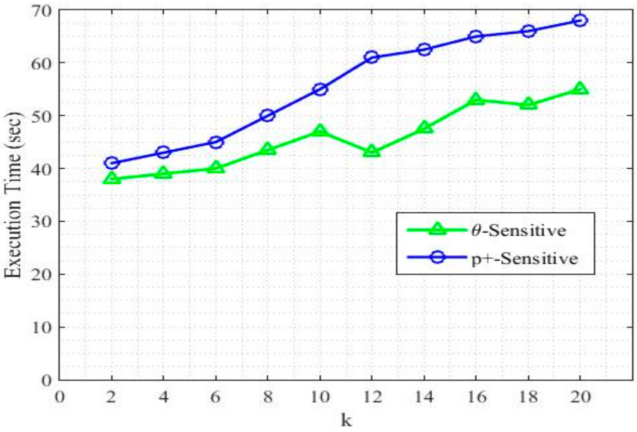

**Figure 7.** Algorithm execution time.

## 7. Conclusions

In this paper, the huge amount of data (i.e., Big Data) collected through the IoT-based devices were anonymized using the proposed $\theta$-sensitive $k$-anonymity privacy model in comparison to $p^+$-sensitive $k$-anonymity model. The purpose was to prevent an attribute disclosure risk in anonymized data. The $p^+$-sensitive $k$-anonymity model was considered to be vulnerable to a privacy breach from sensitive variance, categorical similarity, and homogeneity attacks. These attacks were mitigated by implementing the proposed $\theta$-sensitive $k$-anonymity privacy model using Equation (1). In the proposed solution, the threshold $\theta$ value decides the diversity level for each EC of the dataset. The vulnerabilities in the $p^+$-sensitive $k$-anonymity model and the effectiveness of the proposed $\theta$-sensitive $k$-anonymity model were formally modeled through HLPN, which further ensures the validation of the proposed technique. The experimental work proved the privacy implementation and an improved utility of the released data using different mathematical measures. For future work consideration, the proposed algorithm can be extended to 1:M (single record having many attribute values) [40], to multiple sensitive attributes (MSA) [41–43], or can be modeled by considering the dynamic data set [44] approach.

**Author Contributions:** Conceptualization, R.K. and X.T.; methodology, R.K., X.T., and A.A.; software, R.K.; validation, X.T. and A.A.; formal analysis, R.K., T.K., and S.u.R.M.; writing—original draft preparation, T.K.; writing—review and editing, T.K., X.T., A.A., A.K., W.u.R., and C.M.; supervision, X.T.; funding acquisition, X.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (61932005), and 111 Project of China B16006.

**Conflicts of Interest:** The authors declare that there is no conflict of interests regarding the publication of this paper.

## References

1. Dang, L.M.; Piran, J.; Han, D.; Min, K.; Moon, H. A Survey on Internet of Things and Cloud Computing for Healthcare. *Electronics* **2019**, *8*, 768. [CrossRef]
2. Sun, W.; Cai, Z.; Li, Y.; Liu, F.; Fang, S.; Wang, G. Security and Privacy in the Medical Internet of Things: A Review. *Secur. Commun. Netw.* **2018**, *2018*, 1–9. [CrossRef]
3. Baek, S.; Seo, S.-H.; Kim, S.J. Preserving Patient's Anonymity for Mobile Healthcare System in IoT Environment. *Int. J. Distrib. Sens. Netw.* **2016**, *12*, 2171642. [CrossRef]
4. Liu, F.; Li, T. A Clustering K-Anonymity Privacy-Preserving Method for Wearable IoT Devices. *Secur. Commun. Netw.* **2018**, *2018*, 1–8. [CrossRef]
5. Wan, J.; Al-Awlaqi, M.A.A.H.; Li, M.; O'Grady, M.; Gu, X.; Wang, J.; Cao, N. Wearable IoT enabled real-time health monitoring system. *EURASIP J. Wirel. Commun. Netw.* **2018**, *2018*, 298. [CrossRef]
6. Al-Khafajiy, M.; Baker, T.; Chalmers, C.; Asim, M.; Kolivand, H.; Fahim, M.; Waraich, A. Remote health monitoring of elderly through wearable sensors. *Multimed. Tools Appl.* **2019**, *78*, 24681–24706. [CrossRef]
7. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2002**, *10*, 557–570. [CrossRef]
8. Sweeney, L. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2002**, *10*, 571–588. [CrossRef]
9. Song, F.; Ma, T.; Tian, Y.; Al-Rodhaan, M. A New Method of Privacy Protection: Random k-Anonymous. *IEEE Access* **2019**, *7*, 75434–75445. [CrossRef]
10. Wang, J.; Du, K.; Luo, X.; Li, X. Two privacy-preserving approaches for data publishing with identity reservation. *Knowl. Inf. Syst.* **2018**, *60*, 1039–1080. [CrossRef]
11. Amiri, F.; Yazdani, N.; Shakery, A.; Chinaei, A.H. Hierarchical anonymization algorithms against background knowledge attack in data releasing. *Knowl. Based Syst.* **2016**, *101*, 71–89. [CrossRef]
12. Yaseen, S.; Abbas, S.M.A.; Anjum, A.; Saba, T.; Khan, A.; Malik, S.U.R.; Ahmad, N.; Shahzad, B.; Bashir, A.K. Improved Generalization for Secure Data Publishing. *IEEE Access* **2018**, *6*, 27156–27165. [CrossRef]
13. Liu, X.; Deng, R.H.; Choo, K.K.R.; Weng, J. An efficient privacy preserving outsourced calculation tool kit with multiple keys. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 2401–2414. [CrossRef]
14. Michalas, A. The lord of the shares. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Limassol, Cyprus, 8–12 April 2019; pp. 146–155. [CrossRef]
15. Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. *Int. Conf. Data Eng.* **2006**, *1*, 24. [CrossRef]
16. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy beyond k-Anonymity and l-Diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007; pp. 106–115.
17. Sun, X.; Sun, L.; Wang, H. Extended k-anonymity models against sensitive attribute disclosure. *Comput. Commun.* **2011**, *34*, 526–535. [CrossRef]
18. Anjum, A.; Malik, S.U.R.; Choo, K.-K.R.; Khan, A.; Haroon, A.; Khan, S.; Khan, S.U.; Ahmad, N.; Raza, B. An efficient privacy mechanism for electronic health records. *Comput. Secur.* **2018**, *72*, 196–211. [CrossRef]
19. Campan, A.; Truta, T.M.; Cooper, N. p-sensitive k-anonymity with generalization constraints. *Trans. Data Privacy* **2010**, *3*, 65–89.

20. Al-Khafajiy, M.; Webster, L.; Baker, T.; Waraich, A. Towards fog driven IoT healthcare. In Proceedings of the 2nd International Conference on Future Networks and Distributed Systems, Amman, Jordan, 26–27 June 2018; Volume 9, p. 9.
21. Shahzad, A.; Lee, Y.S.; Lee, M.; Kim, Y.-G.; Xiong, N.N. Real-Time Cloud-Based Health Tracking and Monitoring System in Designed Boundary for Cardiology Patients. *J. Sens.* **2018**, *2018*, 1–15. [CrossRef]
22. Domingo-Ferrer, J.; Soria-Comas, J. From t-closeness to differential privacy and vice versa in data anonymization. *Knowl. Based Syst.* **2015**, *74*, 151–158. [CrossRef]
23. Dwork, C. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12.
24. Fung, B.C.; Wang, K.; Chen, R.; Yu, P.S. Privacy-preserving data publishing. *ACM Comput. Surv.* **2010**, *42*, 1–53. [CrossRef]
25. Xu, Y.; Ma, T.; Tang, M.; Tian, W. A Survey of Privacy Preserving Data Publishing using Generalization and Suppression. *Appl. Math. Inf. Sci.* **2014**, *8*, 1103–1116. [CrossRef]
26. Torra, V. *Transparency in Microaggregation*; UNECE: Skovde, Sweden, 2015; pp. 1–8. Available online: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A861563&dswid=-2982 (accessed on 25 August 2019).
27. Panackal, J.J.; S.Pillai, A. Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets. *Procedia Comput. Sci.* **2015**, *50*, 347–352. [CrossRef]
28. Rahimi, M.; Bateni, M.; Mohammadinejad, H. Extended K-Anonymity Model for Privacy Preserving on Micro Data. *Int. J. Comput. Netw. Inf. Secur.* **2015**, *7*, 42–51. [CrossRef]
29. Sowmiyaa, P.; Tamilarasu, P.; Kavitha, S.; Rekha, A.; Krishna, G.R. Privacy Preservation for Microdata by using k-Anonymity Algorthim. *Int. J. Adv. Res. Comput. Commun. Eng.* **2015**, *4*, 373–375.
30. Wong, C.; Li, J.; Fu, W.; Wang, K. (α,k)-Anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining ACM, Philadelphia, PA, USA, 20–23 August 2006; pp. 754–759.
31. Zhang, Q.; Koudas, N.; Srivastava, D.; Yu, T. Aggregate Query Answering on Anonymized Tables. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Institute of Electrical and Electronics Engineers (IEEE), Istanbul, Turkey, 17–20 April 2007; pp. 116–125.
32. Li, J.; Tao, Y.; Xiao, X. Preservation of proximity privacy in publishing numerical sensitive data. In Proceedings of the 2008 ACM SIGMOD International Conference, Association for Computing Machinery (ACM), Vancouver, BC, Canada, 9–12 June 2008; pp. 473–486. [CrossRef]
33. Xiao, X.; Tao, Y. Personalized privacy preservation. In Proceedings of the 2006 ACM SIGMOD International Conference, Chicago, IL, USA, 27–29 June 2006; p. 229. [CrossRef]
34. Christen, P.; Vatsalan, D.; Fu, Z. Advanced Record Linkage Methods and Privacy Aspects for Population Reconstruction—A Survey and Case Studies. In *Population Reconstruction*; Springer: Berlin, Germany, 2015; pp. 87–110. [CrossRef]
35. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
36. Rubner, Y.; Tomasi, C.; Guibas, L.J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
37. Ali, M.; Malik, S.U.R.; Khan, S.U. DaSCE: Data Security for Cloud Environment with Semi-Trusted Third Party. *IEEE Trans. Cloud Comput.* **2015**, *5*, 642–655. [CrossRef]
38. Bayardo, R.J.; Agrawal, R. Data Privacy through Optimal k-Anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; pp. 217–228.
39. Lefevre, K.; DeWitt, D.; Ramakrishnan, R. Mondrian Multidimensional K-Anonymity. In Proceedings of the 22nd International Conference on Data Engineering, Atlanta, GA, USA, 3–8 April 2006; p. 25.
40. Gong, Q.; Luo, J.; Yang, M.; Ni, W.; Li, X.-B. Anonymizing 1:M microdata with high utility. *Knowl. Based Syst.* **2016**, *115*, 15–26. [CrossRef]
41. Wang, R.; Zhu, Y.; Chen, T.-S.; Chang, C.-C. Privacy-Preserving Algorithms for Multiple Sensitive Attributes Satisfying t-Closeness. *J. Comput. Sci. Technol.* **2018**, *33*, 1231–1242. [CrossRef]
42. Anjum, A.; Ahmad, N.; Malik, S.U.R.; Zubair, S.; Shahzad, B. An efficient approach for publishing microdata for multiple sensitive attributes. *J. Supercomput.* **2018**, *74*, 5127–5155. [CrossRef]

43. Khan, R.; Tao, X.; Anjum, A.; Sajjad, H.; Malik, S.U.R.; Khan, A.; Amiri, F. Privacy Preserving for Multiple Sensitive Attributes against Fingerprint Correlation Attack Satisfying c-Diversity. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 1–18. [CrossRef]
44. Zhu, H.; Liang, H.B.; Zhao, L.; Peng, D.Y.; Xiong, L. τ-Safe (l,k)-Diversity Privacy Model for sequential publication with high utility. *IEEE Access* **2019**, *7*, 687–701. [CrossRef]