An Enhanced and Robust Data Publishing Scheme for Private and Useful 1:M Microdata

Muhammad Rizwan, Ammar Hawbani, Wang Xingfu, Adeel Anjum
Pelin Angin, Yigit Sever, Sanchuan Chen, Liang Zhao, and Ahmed Al-Dubai

Abstract—A data publishing deal conducted with anonymous microdata can preserve the privacy of people. However, anonymizing data with multiple records of an individual (1:M dataset) is still a challenging problem. After anonymizing the 1:M microdata, the vertical correlation can be exploited to launch privacy attacks. In this paper, a novel privacy preserving model l_c, l_s-ANGEL is proposed. To validate the new model, two privacy attacks are presented, namely, a Vertical correlation attack (V_{c0}) and a Vulnerable sensitive attribute attack (V_{sa}) on 1:M datasets, which breach the privacy of individuals. Furthermore, the proposed model is examined through High-Level Petri Nets (HLPNs). Our experiments on three real-world datasets;"INFORMS","YOUTUBE", and "IMDb" demonstrate that the proposed model outperforms the state-of-the-art models. Our practices and lessons learned in this work can direct future concrete steps towards Multiple Sensitive Attributes, where we can expand the proposed model to dynamic datasets.

Index Terms—Internet of Things, big data, electronic health records, privacy of data, k-anonymity, 1:M microdata

1 Introduction

Privacy preserving data publishing techniques focus on sharing a sanitized view of a private dataset to the recipients such as government institutions, research organizations, and statisticians. Private datasets contain sensitive data about individuals. For instance, hospitals release data about patients for research or funding purposes [1], [2], [3], [4]. The data must be processed with certain privacy aspects in mind to preserve an individual's privacy before sharing sensitive data that can include the following features:

- Personally identifiable information (ID, first name, last name etc. – attributes that can uniquely identify an individual);
- Quasi-identifiers (age, gender, material status, nationality, zip code, etc. – the combination of which can identify an individual);
- The private or sensitive attributes (medical history, salary, etc. — must be kept confidential according to individuals' requirements) [3], [5], [6].

The primary challenge of processing the given features is the trade-off between the privacy and the utility of the collected data [7], [8]. The utility of data is maximized when the data is not altered, while data privacy is maximized when the shared data does not resemble the actual data at all. High utility is required in big data analytics, while data privacy is important for the data owners. Therefore, the utility and privacy of data are two highly desirable but incompatible concerns in Privacy-Preserving Data Publication (PPDP) [5], [6]. For example, a previous study [9] revealed that 87% of the US population could be identified using linking attacks that matches *three* quasi-identifiers (QIDs): *gender*, *five-digit zip code* and *date of birth*, using publicly available information like census or voting data. The privacy concerns of data owners are genuine and should be kept in mind to implement privacy measures before publishing the data, otherwise, there will be a loss of trust between the data owner and the data publisher [10], [9]. However, a reasonable balance should be maintained between the data utility and privacy implementation which is an open question in PPDP. Addressing this concern, a plethora of anonymization models; k-anonymity [9], l-diversity [11], t-closeness [12], p,-sensity [13], extended p-sensitivity [14], balanced p^+-sensitive k-anonymity [15] have been proposed (see Section 2 for more details). However, the majority of these privacy models are designed for anonymizing 1:1 records, cases where a person has a single record. A more realistic and practical scenario is the case of 1:M datasets; where a single person has more than one record. Refer to the supplementary material for additional details.

The main contributions of this paper are as follows:

1) A novel privacy-preserving model (l_c, l_s)-ANGEL Algorithm is proposed and we demonstrate its effectiveness against "Vertical correlation attack (V_{co})" and "Vulnerable sensitive attribute attack (V_{sa})" for 1:M data publication.
2) The proposed (l_c, l_s)-ANGEL Algorithm's privacy model is formulated with the state-of-the-art privacy preserving technique 1:M Generalization through High-Level Petri Nets (HLPNs). The formal modelling

Muhammad Rizwan, and Xingfu Wang are with the School of Computer Science and Technology, University of Science and Technology of China Emails:rizwanramay@gmail.com, wangxfu@ustc.edu.cn.
Ammar Hawbani and Liang zhao are with the School of Computer Science, Shenyang Aerospace University. Email: anmande@ustc.edu.cn;lzhao@sau.edu.cn.
Adeel Anjum is with institute of information technology, Quaid I Azam University,Islamabad, Pakistan E-mail:adeelanjum2001@hotmail.com;
Pelin Angin and Yigit Sever are with the Department of Computer Engineering, Middle East Technical University, Turkey ({pangin, yigit}@ceng.metu.edu.tr).
Sanchuan Chen is with the Department of Computer Science and Software Engineering, Auburn University, USA. Email: schen@auburn.edu.
Ahmed Al-Dubai is with the Computing school of Edinburgh Napier University, United Kingdom E-mail: a.al-dubai@napier.ac.uk;
Ammar Hawbani and Xingfu Wang are the corresponding authors.

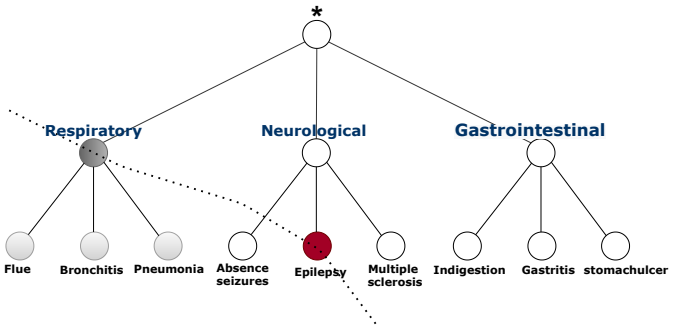


Fig. 1: Lowest common cut on Transaction Generalization Hierarchy

demonstrates the effectiveness of the proposed approach and the invalidation of the 1:M generalization technique for the attacks mentioned above.

3) We provide a detailed analysis and present the privacy-utility trade-off of our proposed approach with experiments on real-world datasets.

The rest of the paper is organized as follows. Section 2 provides a detailed and systematic literature review of the related work. The preliminaries and the problem setting are discussed in Section 3. Section 4 revisits "1:M Generalization" and provides its formal analysis against privacy attacks. The proposed privacy model (l_c, l_s)-ANGEL Algorithm and its formal definition and verification are presented in Subsection 5.1. Section 6 provides experiment results and analyses the results based on privacy-utility evaluation parameters. Section 7 concludes the paper with possible future work directions.

2 Related Work

Privacy aware data publishing involves a trade-off between data privacy and data utility. Researchers proposed anonymization techniques to optimize this trade-off in a balanced manner [14], [16]. One such approach is the use of cryptographic operations. Several privacy techniques that had been built upon cryptographic functions have been proposed [17], [18], [19]. However, such techniques are computationally expensive [20]. In cryptography based techniques, the data cannot be published to a larger and unknown audience as the keys could only be shared with a known audience. Thus, anonymization based privacy models and their underlying techniques perform well in data publication since they are lightweight and easy to use. In the context of health data privacy, various anonymization-based privacy-preserving models and underlying techniques have been proposed.

The main goal of anonymization based privacy models and techniques is to publish individuals' data for the sake of medical and other research or analysis, without compromising individuals' privacy. In the approaches proposed earlier, personally identifiable attributes (e.g. name) were removed from the data before publication. However, linking attacks; matching QIDs with externally available datasets could then be performed. Sweeney [9] was the first to propose a k-anonymity privacy model to prevent linking attacks. The model ensured every record in a particular QIDs-group of the microdata table to be "indistinguishable" from the other $k-1$ records in terms of QIDs. Thus, k-anonymity avoids identity-disclosure attacks, attacks where an adversary identifies a particular individual's identity. Sei et al. proposed models like $(l_1, \ldots, l_q)$-diversity and $(t_1, \ldots, t_q)$-closeness to enhance privacy-preserving data mining by addressing the dual role of sensitive quasi-identifiers. These models provide flexible anonymization methods for attributes serving as both identifiers and sensitive information [21]. Additionally, they introduced the $(l_1, \ldots, l_q)$-diversity model with an anonymization and reconstruction method, evaluated on real datasets [22].

However, the model proposed by Sei et al. [22] cannot protect against attribute disclosure, background knowledge attacks (attacks where an adversary knows enough background knowledge about a particular individual), homogeneous attacks (attacks where all SAs are the same in a particular group), as k-anonymity only deals with QIDs and the adversary could easily get the SA of a particular individual. Another privacy model, l-diversity [11] overcame the attacks mentioned above by ensuring l distinct sensitive attributes (SAs) in all the QIDs-groups of microdata. However, l-diversity cannot protect against similarity attacks, attacks where all l distinct SAs belong to the same genre. For instance, Flu, Chest Infection, and Bronchitis are three distinct diseases, but they all belong to the same genre; Respiratory Tract Infections. In that case, an adversary can unmask the genre of a particular patient. Additional details on privacy techniques and related work are provided in the supplementary file, including a tabular summary.

3 Preliminaries and Problem Setting

In this section, we present the fundamental definitions and notations used in this paper. Moreover, we demonstrate the adversarial model with privacy attack scenarios in the context of the 1:M Generalization privacy technique used in (k, l)-diversity [16].

3.1 Adversary Model

To protect the sensitive records for the data publication, we must assume a potential adversarial model and the possible privacy attacks that the selected model entails. We identified some possible privacy breaches in the context of the 1:M Generalization model;

Scenario 1: The generalized sensitive attributes from the (k, l)-diversity model can be used to identify the current medical status of a patient (e.g. whether their disease is progressing or recovering) using some background knowledge. For instance, Say an adversary has background knowledge about Simon's QIDs (age: 36, sex: male, zip code: 18000) and the fact that he had visited the hospital previously for indigestion treatment, and now he has recently been discharged from the hospital after two days. The details are provided in the supplementary file.

Scenario 2: The lowest common cut in the transactional generalization hierarchy used in (k, l)-diversity model along with some background knowledge can be used to re-identify a particular data owner. For example, if

TABLE 1: Privacy attacks on (k, l)-diversity with corresponding scenarios

Privacy Attacks	Attack Description	Corresponding Scenarios
Vertical Correlation Attacks (V_{co})	An adversary can perform V_{co} attacks to reveal an individual's current medical status (the type and disease's progression or recovery) if they can correlate SAs and background knowledge about the individual	I
Vulnerable Sensitive Attribute Attacks (V_{sa})	An adversary can perform V_{sa} attacks to re-identify an individual if they can successfully correlate the vulnerable SA: left unprotected by lowest common cut on transaction generalization hierarchy, with the background knowledge about that individual	II

an adversary knows the background knowledge about Daisy's QIDs (age: 48, sex: F, zip code: 20000) and the fact that she was admitted to the hospital a few months ago because of her worsening condition due to Pneumonia. The details are provided in the supplementary file.

4 Revisiting 1:M Generalization

In this section, we will review the "1-M Generalization" algorithm's privacy technique in detail. Then, in Subsection 4.2, we will present the formal modelling and analysis of the algorithm to check the correctness and privacy disclosures. The 1:M Generalization is the underlying technique of the (k, l)-diversity privacy model. The said model provides protection for both QIDs and SAs against the privacy leakages through attribute disclosure and linking attacks [16]. As shown in Table 1, privacy attacks on (k, l)-diversity are categorized into vertical correlation attacks (V_{co}) and vulnerable sensitive attribute attacks (V_{sa}). The formal definition of (k, l)-diversity is as follows:

***Definition* 1.** (k, l)-diversity [16]. A 1:M microdata table T satisfies (k, l)-diversity if an only if:

- For any SAFBT (See definition 9 in supplementary file), there exist at-least k individuals.
- For any QIDs-group T, there exist at-least l "well-represented" or "distinct" SA-fingerprints.

4.1 The 1:M Generalization Algorithms

There are three algorithms in 1:M Generalization;

Algorithm 1: Transformation. First perform transformation (See definition 10 in supplementary file)) on a 1:M dataset, then convert it into a 1:1 dataset. The privacy models designed for 1:1 datasets can now be applied.

Algorithm 1: First perform partition. a local recording and a fast top-down anonymization algorithm for set-valued data publication efficiently [23]. The algorithm efficiently partitions the transformed microdata table into k sized QIDs-groups using the SA fingerprints' similarity. The records are anonymized through a transaction generalization technique. After partitioning, a QIDs-group becomes SAFB in which each record is indistinguishable from the other $k-1$ records.

Algorithm 3: Mondrian. First perform `Mondrian` to generalize the QIDs. The `Mondrian` is also a top-down algorithm [24] that can be implemented directly like `Partition`. It anonymizes the QIDs in such a manner that each QIDs-group may satisfy the l-diversity on SAs.

An overview of the 1:M Generalization algorithms is presented at Table 2.

4.2 Formal Modeling and Analysis of Attacks on 1:M Generalization

We demonstrate that our identified privacy attacks - vertical correlation attack (V_{co}) and vulnerable sensitive attribute (V_{sa}) - can occur in 1:M Generalization. Formal verification of 1:M Generalization involves formal modelling of the privacy model through High Level Petri Nets (HLPNs), Z language and an extensive analysis in terms of HLPNs' mathematical properties. The mathematical properties are translated into SMT-Lib and then checked with the Z3 solver to determine whether they hold or not. Details are provided in tabular form in the supplementary file, which shows the mapping of data types in HLPNs.

To reiterate, 1:M Generalization has three steps; `Transformation`, `Partition` and `Mondrian`. `Transformation` maps 1:M records to 1:1 depending on QIDs similarity. It adds different SAs of the same PID to form SA fingerprints whereas PID is also transformed to TID by QIDs similarity. The transformation process is presented in Equation (1).

$$\begin{aligned} R(\text{Transform}) = \forall i2 \in x2, i3 \in x3 \wedge i3[1] \coloneqq i2[1] \\ \wedge\, i3[2] \coloneqq \texttt{Transform}(i2[2], i2[1]) \\ \wedge\, i3[3] \coloneqq \texttt{Transform}(i2[3], i2[1]) \\ \wedge\, x3' \coloneqq x3 \cup \{i3[1], i3[2], i3[3]\} \end{aligned} \tag{1}$$

Equation (2) checks transformed records for whether the condition of k-anonymity for a given input value k is satisfied or not. Then, Equation (3) performs `Generalization` to convert transformed sensitive identifiers (TSI) to Generalized sensitive attributes (GSA).

$$\begin{aligned} R(\text{ChckSplit}) = \forall i4 \in x4, i6 \in x6, i7 \in x7 \\ \vee\, \text{Count}(i4[1]) \geq i5 \rightarrow i7[1] \coloneqq TRUE \\ \wedge\, x7 \coloneqq x7 \cup \{(i7)\} \\ \vee\, \text{Count}(i4[1]) < i5 \rightarrow i7[1] \coloneqq FALSE \\ \vee\, x7 \coloneqq x7 \cup \{(i7)\} \end{aligned} \tag{2}$$

$$\begin{aligned} R(\text{PickNode}) = \forall i8 \in x8, \forall i9 \in x9, \forall i10 \in x10 \\ \vee\, i9 = TRUE \rightarrow i10 \coloneqq \text{Generalize}(i8[2]) \\ \wedge\, x10 \coloneqq x10 \cup \{(i10)\} \end{aligned} \tag{3}$$

$$\begin{aligned} R(\text{DistData}) = \forall i11 \in x11, \forall i12 \in x12, \forall i13 \in x13 \\ \vee\, i13[1] \coloneqq i11[1] \wedge i13[2] \coloneqq Dist-data(i11[2], i10) \\ \wedge\, i13[3] \coloneqq i11[3] \\ \wedge\, x13' \coloneqq x13 \cup \{i13[1], i13[2], i13[3]\} \end{aligned} \tag{4}$$

Algorithm 1: Transformation

Input: $T, k \wedge l$
Output: T

1:M-Generalization(T)
$T'' \leftarrow$ Transform T into 1:1 dataset
$IT \leftarrow$ Partition(T', k)
$T^* \leftarrow$ Modrian(IT, l)
return T

Algorithm 2: Partition for SA fingerprint

```
Partition(partition, k)
if partition cannot be split then
    Add partition to global return list
else
    splitNode ← pick_node(partition)
end
/* pick a node with max.
   information gain                */
subPartitions
  ← distribute_data(partition, splitNode)
/* handle subPartitions with
   < k records                     */
balance_partitions(subPartitions)
for subPartition in subPartitions do
    partition(subPartition)
end
```

Algorithm 3: Modrian for 1:M data

```
Mondrian(partition, l)
if partition cannot be split then
    Add partition to global return list
else
    /* choose attribute with
       the widest values range
       */
    dim ← Choose_Attribute(partition)
end
if dim is numeric then
    threshold ←
      choose_threshold(partition, dim)
    lhs ← {t ∈ partition | t[dim] ≤
      threshold}
    rhs ← {t ∈ partition | t[dim] >
      threshold}
    subPartition ← {lhs} ∪ {rhs}
else
    splitNode ← split(partition, dim)
    subPartitions ←
      distribute_data(partition, splitNode)
end
for subPartition in subPartitions do
    Mondrian(subPartition, l)
end
```

TABLE 2: Overview of 1:M Generalization algorithms

Equation (4) distributes and balances the transformed records to each sub-partition based on GSA fingerprint similarity. Then, Equation (5) balances the GSA of the sub-partition to BSA and stores the record values in place **IDT**. Sub-partitions with less than k records are added to place **G** in Equation (6). Now that both **IDT** and **G** have single record partitions, they are merged into place **CDT** at Equation (7).

$$\begin{aligned} &R(\text{BPartition}) = \forall i14 \in x14, \forall i15 \in x15 \\ &\vee\, i15[1] \coloneqq i14[1] \\ &\wedge\, i15[2] \coloneqq Bal - \text{partition}(i15[3] \coloneqq i14[3]) \\ &\wedge\, x15' \coloneqq x15 \cup \{x15[1], i15[2], i15[3]\} \end{aligned} \tag{5}$$

$$\begin{aligned} R(K') = \forall i16 \in x16, \forall i17 \in x17 \\ \vee\, i16 = FALSE \rightarrow i18[1] \coloneqq i17[1] \\ \wedge\, i18[2] \coloneqq i17[2] \\ \wedge\, i18[3] \coloneqq i17[3] \\ \wedge\, x18' \coloneqq x18 \cup \{i18[1], i18[2], i18[3]\} \end{aligned} \tag{6}$$

$$\begin{aligned} &R(\text{Merge}) = \forall i19 \in x19, \forall i20 \in x20, \forall i21 \in x21 \\ &\vee\, i21[1] \coloneqq i19[1] \wedge i21[2] \coloneqq \text{combine}(i21[3] \coloneqq i19[3]) \\ &\wedge\, x21' \coloneqq x21 \cup \{i21[1], i21[2], i21[3]\} \end{aligned} \tag{7}$$

To anonymize QID values, we start by checking whether the dimension of the QID values is numerical in Equation (8) or categorical in Equation (9). For numerical values of QID, threshold is taken to form, arrange of values for generalized QID in Equation (10) and Equation (11). Equation (12) checks sub-partitions for l different SA fingerprint values. If sub-partitions cannot be split further, then it returns those records to place **G**.

$$\begin{aligned} R(\text{ChoseAttrb}) = \forall i25 \in x25, \forall i26 \in x26 \\ \vee\, i28[1] \coloneqq \text{Chos} - \text{Attrb} \end{aligned} \tag{8}$$

$$\begin{aligned} R(\text{ChckDim}) = \forall i_{27} \in x_{27}, \forall i_{28} \in x_{28} \\ \vee\, i_{28}[1] \coloneqq \text{Chck} - \text{Type} \end{aligned} \tag{9}$$

$$\begin{aligned} R(\text{ChoseThreshold}) = \forall i_{29} \in x_{29}, \forall i_{30} \in x_{30}, \forall i_{31} \in x_{31} \\ \vee\, i_{31}[1] \coloneqq \text{Threshold}(i_{30}[1]) \\ \vee\, x'_{31} \coloneqq x_{18} \cup \{i_{31}[1]\} \end{aligned} \tag{10}$$

$$\begin{aligned} R(\text{ChckThrshld}) = \forall i_{32} \in x_{32}, \forall i_{33} \in x_{33}, \forall i_{34} \in x_{34} \\ \vee\, i_{34}[1] \coloneqq \text{Thrshold}(i_{32}[1], i_{33}[1]) \\ \wedge\, i_{34}[2] \coloneqq \text{Thrshold}(i_{32}[2], i_{33}[2]) \\ \wedge\, x'_{34} \coloneqq x_{34} \cup \{i_{34}[1], i_{34}[2]\} \end{aligned} \tag{11}$$

$$\begin{aligned} &R(\text{Mondrian}) = \forall i_{41} \in x_{41}, \forall i_{42} \in x_{42}, \forall i_{43} \in x_{43} \\ &\vee\, i_{43}[1] = FALSE \rightarrow i_4 4[2] \coloneqq \text{Mondrian}(i_{41}[2], i_{42}[1]) \\ &\wedge\, x'_{44} \coloneqq x_{44} \cup \{i_{44}[1] i_{44}[2], i_{44}[3]\} \end{aligned} \tag{12}$$

SA fingerprint values can be used in V_{co} attacks to re-identify the targeted person using background knowledge and strong posterior belief. Lowest-cut transactional generalization can help the adversary to exploit the vulnerable sensitive attribute in V_{sa} attacks with the assistance of background knowledge and posterior belief. Equation (13) and Equation (14) illustrate V_{co} and V_{sa} attacks, respectively.

$$\begin{aligned} &R(V_{co}\text{Attacks}) = \forall i_{48} \in x_{48}, \forall i_{49} \in x_{49}, \forall i_{50} \in x_{50} \\ &\vee\, V_{co} Dis(i_{48}[2], i_{49}[3]) \rightarrow i_{50}[3] \coloneqq i_{48}[2] \cup i_{49}[3] \coloneqq i_1[3] \end{aligned} \tag{13}$$

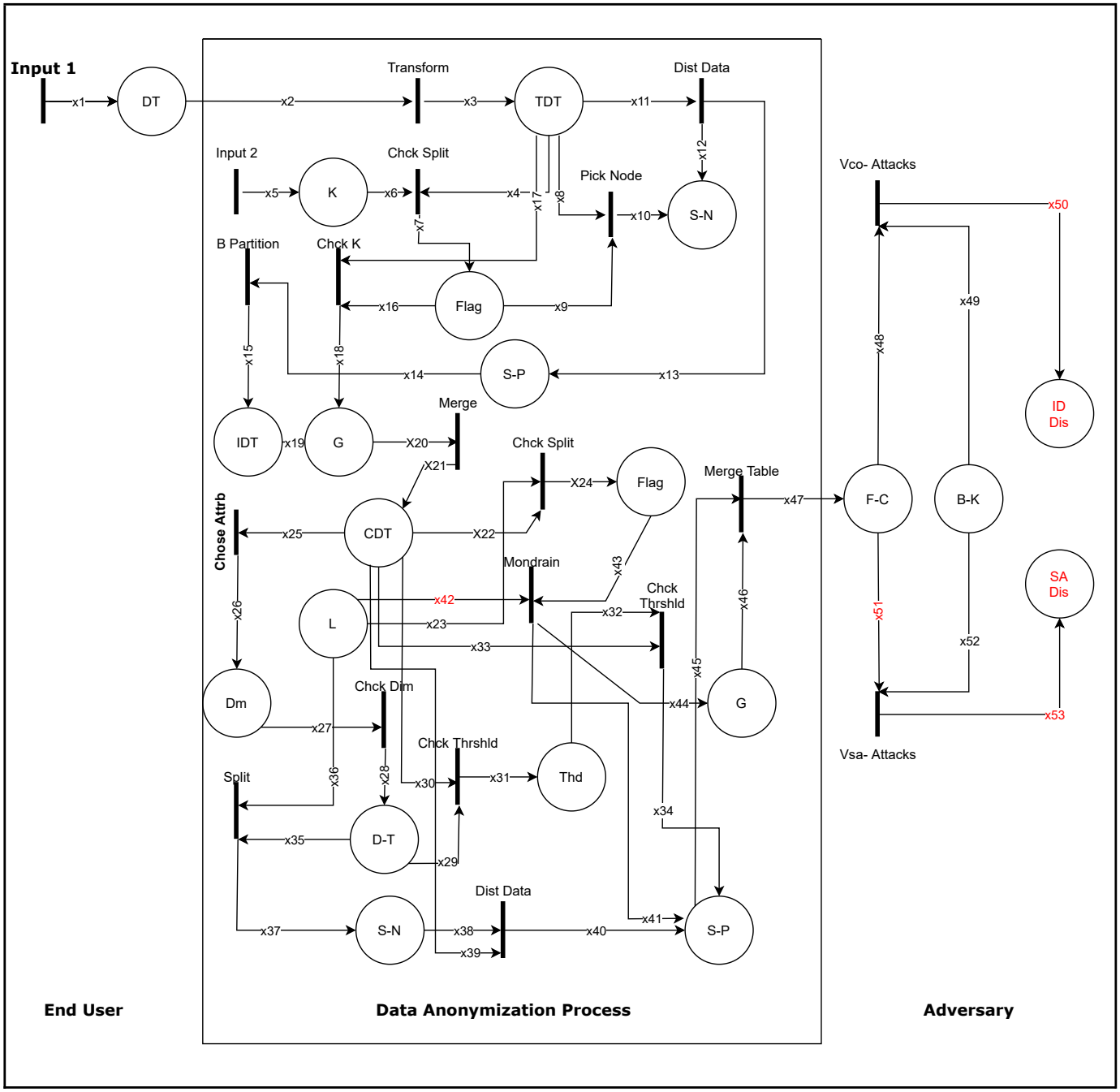


Fig. 2: HLPNs for the identified attacks on 1:M Generalization

$$R(V_{sa}\text{Attacks}) = \forall i_{51} \in x_{51}, \forall i_{52} \in x_{52}, \forall i_{53} \in x_{53} \\ \vee\, V_{sa}Dis(i_{51}[2], i_{52}[3]) \rightarrow i_{53}[2] := (i_{51}[2] \cup i_{52}[3]) := i_1[2] \quad (14)$$

In the V_{co} attack given in Equation (13), the adversary generates an attack using function $V_{co}Dis()$ on the published data with background knowledge and externally available information of an SA value. SA fingerprint values allow an individual to be uniquely identified and SA correlation helps an adversary to perform such attacks. In Equation (14), V_{sa} attacks are performed with function $V_{sa}Dis()$. An adversary also uses the values of SA in the published copy F-C and SAs background knowledge by taking their union in such a way that it allows them to map the original SA value.

5 (l_c, l_s)-ANGEL MODEL

As discussed in Section 2, few approaches have been proposed to deal with the privacy of set-valued or transactional data. The state-of-the-art privacy model (k, l)-diversity with its underlying privacy technique "1:M Generalization" was the earliest to tackle the 1:M privacy issues. In Section 4.2, we demonstrated that the model and technique were vulnerable to vertical correlation and vulnerable sensitive attribute attacks through HLPNs. Moreover, the technique generalized SAs along with QIDs which caused huge information

TABLE 3: Description of Categories and Severities of SAs

SA-Category	SA-Category	Sensitive Attributes (SA)
Respiratory	LOW	Flu
	MILD	Bronchitis
	SEVERE	Pneumonia
Gastrointestinal	LOW	Indigestion
	MILD	Gastritis
	SEVERE	Stomach cancer
Neurological	LOW	Absence Seizures
	MILD	Epilepsy
	SEVERE	Multiple sclerosis

loss. Therefore, we present an improved privacy model, "(l_c, l_s)-ANGEL" to overcome the limitations mentioned above.

(l_c, l_s)-ANGEL is defined as:

***Definition* 2.** (l_c, l_s)-ANGEL. Take a microdata table T, a batch partitioning $A = \{A_1, A_2, \ldots, A_n\}$ and a bucket partitioning $B = \{B_1, B_2, \ldots, B_n\}$. (l_c, l_s)-ANGEL of table T outputs two tables;

- A sensitive table (ST) of the form: $\{S, BID, C\}$ where S represents a sensitive attribute for an individual, BID represents Batch ID and c is the count. For every batch $A_i(1 \leq i \leq n)$ the ST consists of the row (S, i, c) where S is the sensitive attribute and i is the batch ID in such a manner that the sensitive attributes in every batch obeys l_c-diversity on category and l_s-diversity on severity where c is the count.
- A generalized table (GT) is of the form: $\{Q, BID\}$, where Q represents a set of quasi-identifiers attributes for an individual and BID represents the Batch ID. For every bucket $B_i(1 \leq i \leq n)$, the GT consists of the row (Q, i) where Q stores generalized QIDs with i as the batch ID.

The (l_c, l_s)-ANGEL model has four steps; transformation, making l_c-diverse SA-categories, making l_s-diverse SA-severity of l_c-diverse SA-categories, and Angelization. (l_c, l_s)-diversity is used to break the vertical correlation between multiple occurrences of an individual's records. In our identified attack scenarios, we assumed that the disease is either progressing or recovering. Therefore, it was necessary to deal with the categorization and severity of diseases and having their distribution distinct in each QIDs-table. In Table 4 we have merged all sensitive attributes of a QIDs-group together so that the vulnerable SAs attack could be mitigated. The details are discussed in the supplementary file.

5.1 (l_c, l_s)-ANGEL Algorithm

We propose a novel privacy-preserving model, (l_c, l_s)-ANGEL Algorithm for the anonymization of 1:M microdata that has a correlation among multiple records of the same individual. We have assumed an adversarial model where the adversary has background knowledge of the individual for at least one record or can unmask the category of the individual's diseases through personal observations. The (l_c, l_s)-ANGEL algorithm effectively addresses these challenges by employing a sophisticated generalization technique, which ensures the anonymization of sensitive data while maintaining data utility. This approach not only enhances data privacy against potential adversarial attacks but also preserves the integrity and usability of the data for legitimate analytical purposes. The algorithm of our proposed model that prevents such attacks is presented as Algorithm 4. The details are provided into a supplementary file.

Algorithm 4: (l_c, l_s)-ANGEL algorithm

Input: MDT: 1:M Microdata Table
k: Value anonymized QIDs $\in CD$
l_c: value diverse SAs - categories l_s: value diverse SAs - severities
Output: GT: Generalized Table
ST: Sensitive Table

begin Procedure: l_c, l_s-ANGEL
  Let k, l_c, l_s such that $(l_c \wedge k \geq 2) \wedge (l_s = 3)$
  Let C at be a set of SAs - Categories
  Let $S_{ev} = \{S_1, S_2, S_3\}$
  begin Step 1: transformation
    $TMDT \coloneqq$ Transformation(MDT)
    Let $k \subseteq TMDT$
  end
  begin Step 2: Making l_c-diverse SA - categories
    if $k \geq 2$ **then**
      foreach $G_i : \{QI \times SA\} \in k$ **do**
        $CT \coloneqq$ Categorize(SA, Cat)
        $Gi' \coloneqq Gi \cup CT \wedge (\forall CT \in Gi, Gi + 1)$
        $Cn \coloneqq$ ComputeCCount(Distinct(CT))
        if $Cn < l_c$ **then**
          $DCT \coloneqq$ Diverse - $C(CT, l_c)$
        endif
      endfor
    endif
  end
  begin Step 3: Making l_s-diverse SA - severity
    $SvT \coloneqq$ Severitise(DCT, Sev)
    $Cs \coloneqq$ ComputeScount(Distict(SvT))
    if $Cs \neq l_s$ **then**
      $DST \coloneqq$ Diverse $- S(SvT, l_s)$
    endif
  end
  begin Step 4: Angelization
    $GT \coloneqq$ Angelize(DST)
    $ST \coloneqq$ Angelize(DST)
    return GT, ST
  end
end

5.2 Formal Modeling and Analysis of l_c, l_s-ANGEL Algorithm using HLPNs

Now, we present the formal definition and analysis of our model through HLPNs and then verify it against our identified attacks. The details are provided in tabular form in the supplementary file. The symbols used in HLPNs of our model and mapping of data types on places respectively. Figure 2 depicts the formal verification through HLPNs. In

TABLE 4: l_c, l_s-diverse table of 1:M Microdata

Personally Identifiable Attributes		Quasi-Identifiers Attributes (QID)			Sensitive Attributes (SA)		
Tuple ID	PID	Age	Sex	Zip Code	Disease	Category	Severity
1-4(Bob)	1	21	M	12000	Flu Bronchitis Pneumonia Epilepsy	Respiratory Respiratory Respiratory Neurological	LOW MILD HIGH MILD
6(Sara)	3	24	F	14000	Epilepsy	Neurological	MILD
12-13(Alice)	6	28	F	11000	Gastritis Stomach Ulcer	Gastrointestinal Gastrointestinal	MILD HIGH
5(David)	2	32	M	17000	Multiple sclerosis	Neurological	HIGH
7-8(Simon)	4	36	M	18000	Indigestion Gastritis	Gastrointestinal Gastrointestinal	LOW MILD
9-11(Daisy)	5	40	F	20000	Flu Pneumonia Epilepsy	Respiratory Respiratory Respiratory	LOW HIGH MILD

TABLE 5: Comparison between the Proposed Algorithm and the Algorithm in [4]

Comparison Criteria	**Proposed l_c, l_s-ANGEL Model**	**(p, l) Angelization Model (Reference [4])**
Objective	Address Vco and Vsa attacks, ensure privacy in 1:M datasets	Mitigate Scor, Nmcor, and Qcor attacks in multi-sensitive datasets
Techniques	1:M Generalization, HLPNs for formal verification	Batch and sensitive batch partitioning for generalization
Privacy Goals	Mitigates vertical correlation and protects vulnerable attributes	Prevents specific correlation attacks (Scor, Nmcor, Qcor)
Data Utility	Reduces information through targeted SA generalization	Balances utility with privacy via sensitive batch tables
Experimental Validation	Tested on real-world datasets (INFORMS, YOUTUBE, IMDb)	Evaluated on synthetic datasets for scalable partitioning
Scalability	Focus on static datasets with future work towards Big Data scalability	Highly scalable for large, static datasets
Future Adaptability	Designed for extension to dynamic datasets and Big Data applications	Primarily applicable to current static dataset structures

Equation (15) the 1:M records from place MDT are transformed to 1:1 depending on the QIDs similarity. Different SAs with the same PID form a SA fingerprint whereas PID is also transformed to TID by QIDs similarity in transition `Transform`. In Equation (16), first we categorize the SAs into appropriate categories with `Catdisease()`, then in Equation (17), l_c value is computed and condition is checked on the SAs-category. If condition is false, then SAs with category is swapped to make the group of SAs and category l-diverse depending upon l_c value. `DiverseC()` makes the group l_c-diverse and stores it in place of DCT in Equation (18). The (l_c, l_s)-ANGEL model in this paper targets vertical correlation and vulnerable sensitive attribute attacks in 1:M datasets, using 1:M Generalization and formal verification through **High-Level Petri Nets (HLPNs)**. Unlike the (p, l)-Angelization algorithm in[4], which addresses correlation attacks via batch partitioning, (l_c, l_s)-ANGEL provides robust privacy with reduced information loss on real-world datasets. While both models achieve privacy preservation, (l_c, l_s)-ANGEL is poised for future extensions to handle dynamic data and Big Data applications, positioning it as a scalable solution for complex environments.

$$\begin{aligned} R(\text{Transform}) = \forall i_2 \in x_2, i_3 \in x_3 \\ \vee\, i_3[3] \coloneqq i_2[3] \\ \wedge\, i_3[1] \coloneqq \text{Transf}(i_2[1], i_2[3]) \\ \wedge\, i_3[2] \coloneqq \text{Transf}(i_2[2], i_2[3]) \\ \wedge\, x_3' \coloneqq x_3 \cup \{i_3[1], i_3[2], i_3[3]\} \end{aligned} \tag{15}$$

$$\begin{aligned} R(\text{Categorize}) = \forall i_4 \in x_4, i_5 \in x_5, i_6 \in x_6 \\ \vee\, i_6[1] \coloneqq \text{Catdisease}(i_4[2], i_5[1]) \\ \wedge\, i_6[2] \coloneqq \text{Catdisease}(i_4[2], i_5[1]) \\ \wedge\, i_6 \coloneqq x_6 \cup \{(i_6[1], i_6[2])\} \end{aligned} \tag{16}$$

$$\begin{aligned} R(\text{ComputeCCount}) = \forall i_7 \in x_7, \forall i_8 \in x_8 \\ \vee\, i_8[2] \coloneqq \text{Count}(\text{Distinct}(i_7[2])) \\ \wedge\, x_8 \coloneqq x_8 \cup \{i_8[2]\} \end{aligned} \tag{17}$$

$$\begin{aligned} R(\text{Diverse} - C) = \forall i_9 \in x_9, \forall i_{10} \in x_{10}, \forall i_{11} \in x_{11} \\ \vee\, i_{10}[1] > i_{10}[2] = FALSE \\ \rightarrow ((i_{11}[2], i_{11}[3]) \coloneqq \text{DiverseC}(i_9[1], i_{10}[2])) \\ \wedge\, x_{11} \coloneqq x_{11} \cup \{(i_{11}[2], i_{11}[3])\} \end{aligned} \tag{18}$$

$$\begin{aligned} R(\text{Severities}) = \forall i_{12} \in x_{12}, \forall i_{13} \in x_{13}, \forall i_{14} \in x_{14} \\ \vee\, i_{14}[1] \coloneqq i_{12}[1] \\ \wedge\, (i_{14}[2], i_{14}[4]) = \text{Sev} - \text{allotment}(i_{12}[2], i_{13}[5]) \\ \wedge\, i_{14}[3] \coloneqq i_{12}[3] \\ \wedge\, x_{14} \coloneqq x_{14} \cup \{(i_{14}[1], i_{14}[2], i_{14}[3], i_{14}[4])\} \end{aligned} \tag{19}$$

$$R(\text{ComputeSCount}) = \forall i_{15} \in x_{15}, \forall i_{16} \in x_{16} \vee i_6[4] \coloneqq \text{Count}(\text{Distinct}(i_5[5])) \wedge x_{16} \coloneqq x_{16} \cup (i_{16}[4]) \quad (20)$$

In transition `Severities` we classify the SAs of every category according to their severity and the output tuples are stored in SvT (Equation (19)). Then function `Count()` computes the count of distinct category severity l_s and the diversity condition on the category severity is checked. If the condition is false then `DiverseS()` makes the group l_s-diverse based on category severity and stored in DST Equations (20),21). During the Angelize transition we split DST into two different tables: a Generalized Table (GT) and a Sensitive Table (ST). Both tables are stored in BDT. This transition is applied to break any correlation between the QIDs and SAs values (Equation (22)). The last transitions present the V_{co} and V_{sa} attacks that we have shown with the discussion on 1:M Generalization in Subsection 4.2. In the proposed algorithm (l_c, l_s)-ANGEL verification, transitions 23, 24 show that privacy attacks V_{co} and V_{sa} are effectively mitigated due to l_c-diverse SAs categories, l_s-diverse SAs severities, and finally with the Angelization technique outputting separate GT and ST tables to break any link between QIDs and SAs.

$$\begin{aligned} R(\text{Diverse} - S) = \forall i_{17} \in x_{17}, \forall i_{18} \in x_{18}, \forall i_{19} \in x_{19} \\ \vee i_{18}[4] > i_{18}[3] = FALSE \\ \rightarrow (i_{19}[3], i_{19}[4]) \coloneqq \text{DiverseS}(i_{17}[3], i_{17}[4]) \\ \wedge i_{19}[1] \coloneqq i_{17}[1] \\ \wedge i_{19}[2] \coloneqq i_{17}[2] \\ \wedge x_{19} \coloneqq x_{19} \cup \{(i_{19}[1], i_{19}[2], i_{19}[3], i_{19}[4])\} \end{aligned} \quad (21)$$

$$\begin{aligned} R(\text{Angelize}) = \forall i_{20} \in x_{20}, \forall i_{21} \in x_{21} \\ \vee i_{21}[1] \coloneqq \text{Angle} \\ \wedge (i_{21}[2]_n \forall i_{21}[2]_n \in x_{21}) \\ \coloneqq \text{Angle}(i_{20}[2], i_{20}[3], i_{20}[4]) \\ \wedge x_{22} \coloneqq \cup\{(i_{22}[1], i_{22}[2])\} \end{aligned} \quad (22)$$

$$\begin{aligned} R(V_{co}\text{Attacks}) = \forall i_{22} \in x_{22}, \forall i_{23} \in x_{23}, \forall i_{24} \in x_{24} \\ \vee V_{co}\text{Dis}(i_{22}[2], i_{23}[3]) \neq i_{24}[3] \\ \wedge i_{22}[2] \cap \imath_{23}[3] = \phi \end{aligned} \quad (23)$$

$$\begin{aligned} R(V_{sa}\text{Attacks}) = \forall i_{25} \in x_{25}, \forall i_{26} \in x_{26}, \forall i_{27} \in x_{27} \\ \vee V_{sa}\text{Dis}(i_{25}[2], i_{26}[3]) \neq i_{27}[2] \\ \wedge (i_{25}[2] \cap i_{26}[3]) = \phi \end{aligned} \quad (24)$$

6 Experimental Evaluation

In this section, we present the experimental evaluation of our proposed model and illustrate the performance comparisons with 1:M generalization [16], and (p, l)-Angelization[4] We have implemented the proposed model and evaluated it on real-world datasets in terms of data utility, privacy and computational efficiency. In Subsection 6.1, we will present the experiment preparation and settings we have used. The datasets and the evaluation parameters as well as the measures for data utility and computational efficiency are also presented here. Subsection 6.2 provides an extensive discussion and analysis on the results obtained through Normalized Certainty Penalty (NCP). Subsection 6.3 presents the query accuracy investigation and Subsection 6.4 shows the comparisons of execution runtimes for 1:M Generalization, (p, l)- Angelization and the proposed model (l_c, l_s)-ANGEL on the datasets.

6.1 Preparation and Setting

We have implemented 1:M Generalization and (l_c, l_s)-ANGEL in Python. All experiments have been conducted on a computer with an 8th generation Intel Core i7 processor, 8 GB RAM and Windows 10. We used three real-world datasets, `INFORMS`, `YOUTUBE` and `IMDB`. All of these datasets are well-known and have been widely used in numerous related works [16], [25], [26], [27].

See Table 6 for an overview of the datasets. The details are provided in the supplementary file.

6.2 Normalized Certainty Penalty

We used NCP [16], [28] to evaluate the information loss of our proposed model. The obtained results have been compared and critically analysed with NCP in 1:M Generalization. NCP evaluates the level of accuracy for all equivalence classes and assigns penalties for information loss caused by the generalization process. Given a microdata table take the value of an attribute A as v. Then, the basic formula for measuring NCP is as follows:

$$\text{NCP}(v) = \begin{cases} 0 & |v| = 1 \\ \frac{|v|}{|A|} & \text{otherwise} \end{cases} \quad (25)$$

In Equation (25), $|v|$ represents the total number of leaf-nodes covered in value v according to generalization hierarchies while $|A|$ is the total number of leaf-nodes in attribute A [16]. Gong et. al. have used two types of NCPs to measure information loss on both QIDs and SAs in 1:M Generalization [12]. For example, if the "epilepsy" in record 1 in the supplementary file is generalized to "Neurological", the information loss would be:

$$\text{NCP}(\text{Neurological}) = \frac{2}{6} = \frac{1}{3} = 0.33$$

Similarly, if the age value 16 is generalized to [16-20] and we assume age range [1, 100] and zip-code range [10001-30000], same as assumed in [12], then the information loss would be:

$$\text{NCP}([16 - 20]) = \frac{5}{100} = 5\%$$

The Equation (25) represents information loss from generalizing an attribute. For a whole table T, the information loss on both QIDs and SAs generalization is calculated using Equation (26) and (27) respectively.

$$\text{QIDs} - \text{NCP}(T) = \frac{\sum_{i=1}^{n} \text{QIDs} - \text{NCP}(r_i)}{n} \quad (26)$$

TABLE 6: Descriptions of the datasets used

Dataset	n	QIDs	SAs	SA Domain
IN-FORMS	58.5 68	Month of Birth Year of Birth Race Years of Education Income	Diagnosis Code	632
YOUTUBE	85.6 07	Age Rate Ratings	Related Videos	117.752
IMDb	1000	Year Genre Director	Rating Gross	7.95 21

Here, T represents the microdata table that has r records in it. r^i is the i^{th} record in T. The QIDs-NCP [16] for table T would be the aggregated QIDs-NCP for all the records of T over total number of records n. For aggregation, QIDs-NCP for a record r can be calculated as follows in Equation (27).

$$\text{QIDs} - \text{NCP}(r) = \frac{\sum_{j=1}^{d} \text{NCP}(r.q_j)}{d} \tag{27}$$

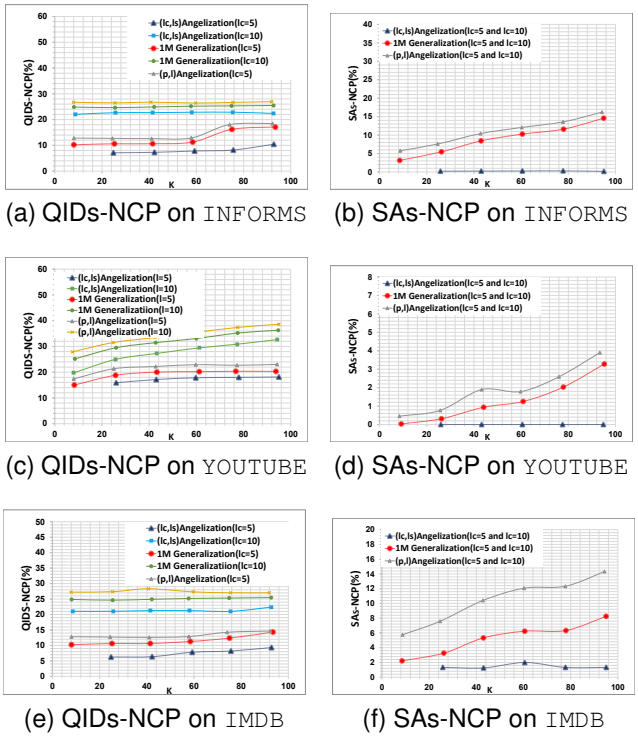
In the above equation, the values of QIDs for record r would be $r.q_1, r.q_2, \ldots, r.q_d$. The QIDs-NCP for record r would be aggregated NCP on all QIDs values of r. For instance, the QIDs-NCP of details are provided in tabular form in the supplementary file:

$$\frac{\left[\{3 \times \frac{10}{100} + \frac{5000}{30000}\} + \{3 \times \frac{10}{100} + \frac{5000}{30000}\}\right]}{6} = 15.6\%$$

On the other hand, the SAs-NCP[16] for the table T can be calculated by the Equation (28).

$$\text{SA} - \text{NCP}(T) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{C(r_i[d+1])} \text{NCP}(j)}{\sum_{i=1}^{n} C\left(r_i[d+1]\right)} \tag{28}$$

Here, $C\left(r_i[d+1]\right)$ represents the number of distinct SAs in SA-fingerprints for record r_i and s be a SA value in $r_i[d+1]$. 1:M Generalization has generalized the sensitive attributes into SA-fingerprints. The generalization of SAs caused huge information loss for sensitive attributes accuracy. Therefore, we have not applied any sort of anonymization on SAs in our proposed model; (l_c, l_s)-ANGEL. That is the main reason that the SAs-NCP on anonymized datasets by our model is zero, regardless of varying k, l, and n parameters on both datasets. In Figure 3, we have plotted results of QIDs-NCP and SA-NCP of 1:M Generalization and (l_c, l_s)-ANGEL with varying k on all datasets; `INFORMS`, `YOUTUBE` and `IMDB`. We have taken different values k on x-axis and QIDs-NCP and SAs-NCP on y-axis. The figure shows that more utility was preserved on both QIDs and SAs by (l_c, l_s)-ANGEL. In our model we have incorporated Angelization [29]; which is a utility preserving privacy technique and l-diversity on SAs' categories and their underlying severities gave well represented values but with



(a) QIDs-NCP on `INFORMS` (b) SAs-NCP on `INFORMS`

(c) QIDs-NCP on `YOUTUBE` (d) SAs-NCP on `YOUTUBE`

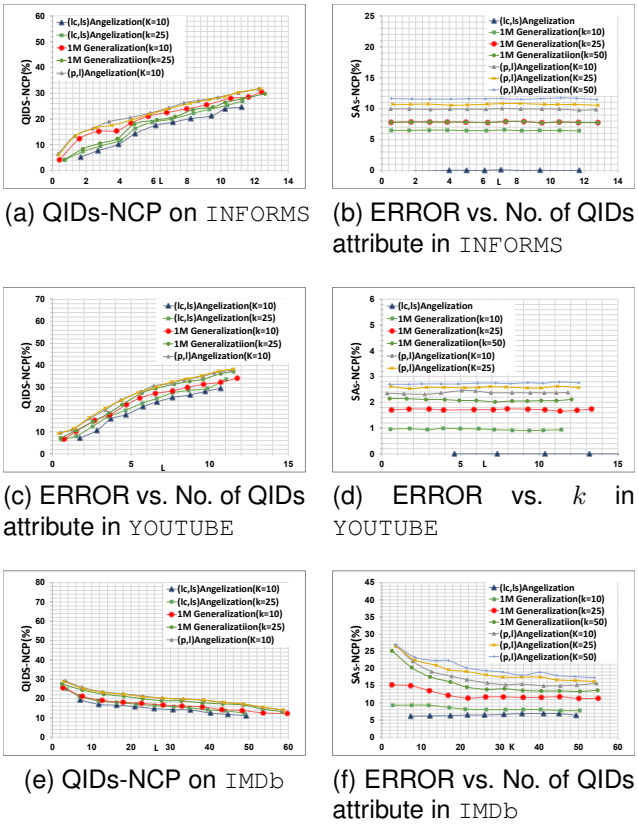
(e) QIDs-NCP on `IMDB` (f) SAs-NCP on `IMDB`

Fig. 3: Information loss with varying values of k

higher utility. The results have been calculated with $l = 5$ and $l = 10$. In Figures 3a and 3c, both 1:M Generalization and (l_c, l_s)-ANGEL has shown a gradual increase in QIDs-NCP on increasing l size. The reason is; a larger l needs to produce larger QIDs groups and consequently, greater QIDs-NCP. We have used different values for l_c while $l_s = 3$ as we have assumed three severities/intensities for our model. It is to be noted that both QIDs-NCP and SAs-NCP in 1:M Generalization increase when k is increased. The reason is the nature of the partition algorithm. The Partition algorithm distorts more information in achieving the k-anonymity on SA-fingerprints. As Mondrian does not change SAs and partition is unrelated to l, therefore the both SAs-NCP with $l = 5$ and $l = 10$ are same and equal in Figures 3b and 3d.

Similarly, Figure 5 depicts the results of QIDs-NCP and SA-NCP of both approaches with varying l on `INFORMS`, `YOUTUBE` and `IMDB`. We have taken different values l on the x-axis and QIDs-NCP and SAs-NCP on the y-axis. Due to almost the same reasons mentioned above, (l_c, l_s)-ANGEL has preserved more utility on both QIDs and SAs than 1:M Generalization. To evaluate data utility in terms of total number of records in the dataset (n), we have followed the same sample-sets formation steps carried out in evaluation of 1:M Generalization.

The Figure 5 shows the results of QIDs-NCP and SAs-NCP of both approaches with varying n both datasets. We have taken a number of records n on the x-axis and QIDs-NCP and SAs-NCP on the y-axis. It can be observed from the figure that (l_c, l_s)-ANGEL has preserved more utility on both QIDs and SAs as compared to 1:M Generalization.In QIDs-NCP evaluation, NCP for both approaches raised with the different sizes of k. In 1:M Generalization, both QIDs-

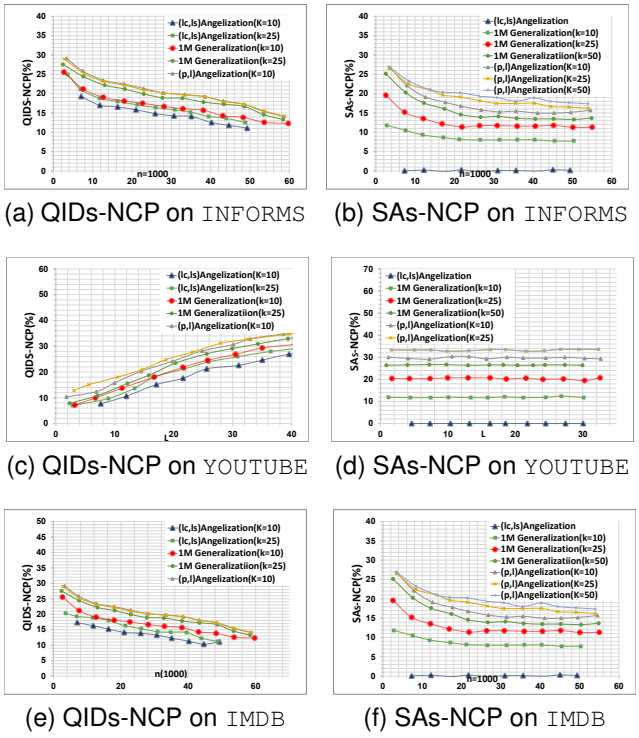


(a) QIDs-NCP on INFORMS (b) ERROR vs. No. of QIDs attribute in INFORMS

(c) ERROR vs. No. of QIDs attribute in YOUTUBE (d) ERROR vs. k in YOUTUBE

(e) QIDs-NCP on IMDb (f) ERROR vs. No. of QIDs attribute in IMDb

Fig. 4: Query Errors



(a) QIDs-NCP on INFORMS (b) SAs-NCP on INFORMS

(c) QIDs-NCP on YOUTUBE (d) SAs-NCP on YOUTUBE

(e) QIDs-NCP on IMDB (f) SAs-NCP on IMDB

Fig. 5: Information loss with varying values of n

NCP and SAs-NCP reduce as the dataset grows because of the sensitivity of both partition and Mondrian to the size of datasets.

6.3 Query Accuracy

Apart from NCP, the data utility of the anonymized or published datasets is measured in query accuracy as well. The estimation is carried out by answering the aggregated queries throughout the datasets. "COUNT" is used to respond to aggregated queries. The QIDs are regarded as the query-predicates. If Q represents the set of quasi-identifiers as, $Q = \{q_1, q_2, \ldots, q_n\}$ and the domain of quasi-identifiers as $D(q_i)$, then the aggregated query can be written as in Equation (29) (30), (31) below.

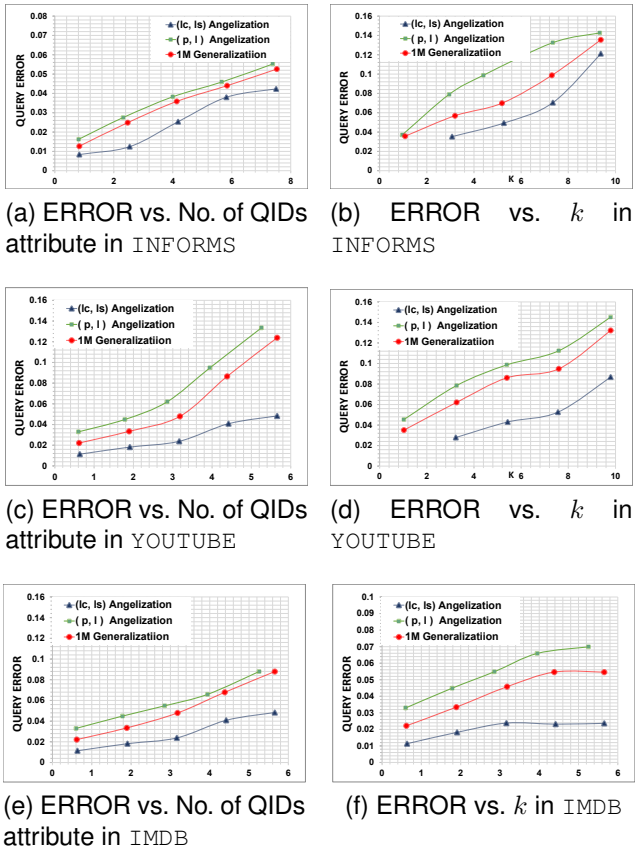
$$\begin{aligned} &\text{SQL Query} = \\ &\text{SELECTCOUNT}() \text{ table } T \\ &\text{WHERE } q_i \in D(q_i) \wedge \ldots \wedge q_n \in D(q_n) \end{aligned} \tag{29}$$

The query predicate holds two major parameters:

(i): q: the query-dimensionality
(ii): θ: the query selectivity

The query-dimensionality parameter is specified by the number of QIDs used in the predicate while the query-selectivity represents the number of values for each attribute. The query-selectivity is calculated as in Equation (30).

$$\theta = \frac{|T_Q|}{|T|} \tag{30}$$

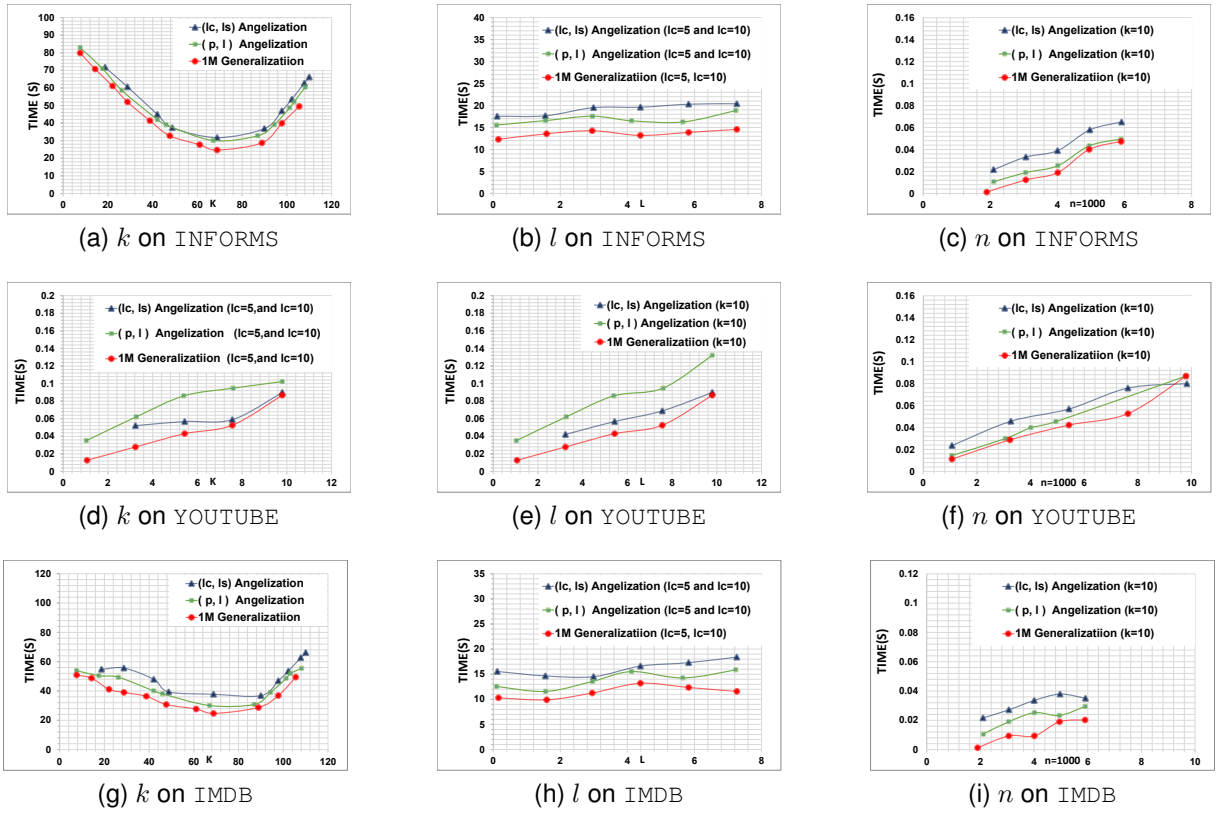


(a) ERROR vs. No. of QIDs attribute in INFORMS (b) ERROR vs. k in INFORMS

(c) ERROR vs. No. of QIDs attribute in YOUTUBE (d) ERROR vs. k in YOUTUBE

(e) ERROR vs. No. of QIDs attribute in IMDB (f) ERROR vs. k in IMDB

Fig. 6: Query Errors



(a) k on INFORMS (b) l on INFORMS (c) n on INFORMS

(d) k on YOUTUBE (e) l on YOUTUBE (f) n on YOUTUBE

(g) k on IMDB (h) l on IMDB (i) n on IMDB

Fig. 7: Execution time for different sizes of k, l, and n on different datasets

The $|T|$ represents the count of tuples in the datasets while the result-set generated from query Q on table T includes the number of tuples. The result of query error is represented by Relative Error; Error(R). The Relative error is a normalized-divergence between the resultant dataset. The relative error is calculated using Equation (31).

$$\text{Relative Error} = \frac{(\text{Estimated Value} - \text{Actual Value})}{\text{Actual Value}} \quad (31)$$

In the formula above, the actual value is the result obtained from the COUNT query on the raw dataset and the estimated count results from the count queries. The Figure 6 shows the query errors results obtained from both approaches over INFORMS, YOUTUBE and IMDB. We have calculated the query error on two parameters; different sizes of k and number of QIDs. The query errors have been taken on the y-axis and two parameters have been taken on the x-axis. The results show that 1:M Generalization has greater query errors than our proposed model (l_c, l_s)-ANGEL on all parameters and datasets. The reason is; 1:M generalization has generalized the sensitive attributes; $a_1, a_2 >$ to A. However, our approach has not generalized the sensitive attributes, rather published them as they are, in the angelized sensitive table.

6.4 Execution Time

The computational efficiency of a model is expressed in terms of its total execution time, and it is considered a significant parameter to evaluate the computational efficiency of an algorithm.

The execution time is calculated with respect to several parameters; size of k, l or n etc. Several research works [30], [26] have used execution time in the context of privacy preserving data publication for evaluating the computational efficiency of proposed models. We have calculated and compared the execution time of both 1:M Generalization and (l_c, l_s)-ANGEL on different values of k and l (in our case $l = l_c$). Figure 7 shows the results obtained in this regard. 1:M Generalization includes three steps while (l_c, l_s)-ANGEL is a four step algorithm. (l_c, l_s)-ANGEL also performs Angelization to break any possible connection between QIDs and SAs so that linking attacks can be mitigated. Therefore, the execution time of our proposed model is higher than 1:M Generalization's execution time in terms of seconds. However, considering the results of both approaches in terms of preserving privacy and utility, the execution time difference is negligible. Figure 7a shows the execution time results obtained on varying sizes of k, l and n on datasets INFORMS, YOUTUBE and IMDB.

7 Conclusion and Future Directions

In this paper, we introduced a novel privacy-preserving model l_c, l_s-ANGEL. The proposed model mitigates the vertical correlation and vulnerable sensitive attribute privacy attacks with a state-of-the-art 1:M Generalization technique. Furthermore, we identified and reported on the information loss due to the generalization of sensitive attributes in the previous models. The main contribution of the paper has been the formal modelling, verification and analysis of the

proposed and prior privacy models in the context of the privacy attacks we have identified. We have also tested the proposed algorithm on datasets commonly used by the literature for a fair comparison. The algorithm we have proposed has been tested on static datasets. However, the work is poised for extension to handle Multiple Sensitive Attributes and is particularly adaptable for Big Data application scenarios. For future work, we aim to enhance the scalability and efficiency, making it suitable for dynamic and large-scale data environments. This adaptation is a critical step towards addressing the complexities and challenges inherent in Big Data applications, thus, making a concrete step towards a number of interesting research directions

8 Acknowledgement

This work was supported in part by the Open Fund of Anhui Engineering Research Centre for Intelligent Applications and Security of Industrial Internet, under Grant IASII24-04.

References

[1] L. M. Dang, M. J. Piran, D. Han, K. Min, and H. Moon, "A Survey on Internet of Things and Cloud Computing for Healthcare," *Electronics*, vol. 8, no. 7, p. 768, Jul. 2019.

[2] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Security and Privacy in the Medical Internet of Things: A Review," *Security and Communication Networks*, vol. 2018, p. e5978636, Mar. 2018.

[3] S. Baek, S.-H. Seo, and S. Kim, "Preserving Patient's Anonymity for Mobile Healthcare System in IoT Environment," *International Journal of Distributed Sensor Networks*, vol. 12, no. 7, p. 2171642, Jul. 2016.

[4] T. Kanwal, S. A. A. Shaukat, A. Anjum, K.-K. R. Choo, A. Khan, N. Ahmad, M. Ahmad, S. U. Khan *et al.*, "Privacy-preserving model and generalization correlation attacks for 1: M data with multiple sensitive attributes," *Information Sciences*, vol. 488, pp. 238–256, 2019.

[5] F. Liu and T. Li, "A Clustering K-Anonymity Privacy-Preserving Method for Wearable IoT Devices," *Security and Communication Networks*, vol. 2018, Jan. 2018.

[6] M. A. Azad, J. Arshad, S. Mahmoud, K. Salah, and M. Imran, "A privacy-preserving framework for smart context-aware healthcare applications," *Transactions on Emerging Telecommunications Technologies*, vol. n/a, no. n/a, May 2019.

[7] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, Jun. 2013.

[8] R. Khan, X. Tao, A. Anjum, H. Sajjad, S. u. R. Malik, A. Khan, and F. Amiri, "Privacy preserving for multiple sensitive attributes against fingerprint correlation attack satisfying c-diversity," *Wireless Communications and Mobile Computing*, vol. 2020, pp. 1–18, 2020.

[9] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, Oct. 2002.

[10] G. Duncan and D. Lambert, "The Risk of Disclosure for Microdata," *Journal of Business & Economic Statistics*, vol. 7, no. 2, pp. 207–217, 1989.

[11] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[12] N. Li, T. Li, and S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, Apr. 2007, pp. 106–115.

[13] A. Campan, T. M. Truta, and N. Cooper, "P-Sensitive K-Anonymity with Generalization Constraints," p. 25, 2010.

[14] X. Sun, H. Wang, J. Li, and T. M. Truta, "Enhanced P -Sensitive K-Anonymity Models for Privacy Preserving Data Publishing," *Trans. Data Privacy*, p. 14, 2008.

[15] R. Khan, X. Tao, A. Anjum, T. Kanwal, S. U. R. Malik, A. Khan, W. U. Rehman, and C. Maple, "θ-sensitive k-anonymity: An anonymization model for iot based electronic health records," *Electronics*, vol. 9, no. 5, p. 716, 2020.

[16] Q. Gong, J. Luo, M. Yang, W. Ni, and X.-B. Li, "Anonymizing 1:M microdata with high utility," *Knowledge-Based Systems*, vol. 115, pp. 15–26, Jan. 2017.

[17] D. He, N. Kumar, H. Wang, L. Wang, K.-K. R. Choo, and A. Vinel, "A Provably-Secure Cross-Domain Handshake Scheme with Symptoms-Matching for Mobile Healthcare Social Network," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 633–645, Jul. 2018.

[18] X. Liu, R. H. Deng, K.-K. R. Choo, and J. Weng, "An Efficient Privacy-Preserving Outsourced Calculation Toolkit With Multiple Keys," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 11, pp. 2401–2414, Nov. 2016.

[19] M. Joshi, K. Joshi, and T. Finin, "Attribute Based Encryption for Secure Access to Cloud Based EHR Systems," in *2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*, Jul. 2018, pp. 932–935.

[20] F. Xhafa, J. Li, G. Zhao, J. Li, X. Chen, and D. S. Wong, "Designing cloud-based electronic health record system with attribute-based encryption," *Multimedia Tools and Applications*, vol. 74, no. 10, pp. 3441–3458, May 2015.

[21] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness," *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 4, pp. 580–593, 2019.

[22] Y. Sei, T. Takenouchi, and A. Ohsuga, "(l1, ..., lq)-diversity for anonymizing sensitive quasi-identifiers," in *2015 IEEE Trustcom/BigDataSE/ISPA*, vol. 1, 2015, pp. 596–603.

[23] Y. He and J. F. Naughton, "Anonymization of set-valued data via top-down, local generalization," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 934–945, Aug. 2009.

[24] 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, pp. 1–1, Jan. 2013.

[25] G. Poulis, G. Loukides, A. Gkoulalas-Divanis, and S. Skiadopoulos, "Anonymizing Data with Relational and Transaction Attributes," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Eds. Berlin, Heidelberg: Springer, 2013, pp. 353–369.

[26] T. Kanwal, S. A. A. Shaukat, A. Anjum, S. u. R. Malik, K.-K. R. Choo, A. Khan, N. Ahmad, M. Ahmad, and S. U. Khan, "Privacy-preserving model and generalization correlation attacks for 1:M data with multiple sensitive attributes," *Information Sciences*, vol. 488, pp. 238–256, Jul. 2019.

[27] G. Loukides, J. Liagouris, A. Gkoulalas-Divanis, and M. Terrovitis, "Disassociation for electronic health record privacy," *Journal of Biomedical Informatics*, vol. 50, pp. 46–61, Aug. 2014.

[28] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization using local recoding," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: Association for Computing Machinery, Aug. 2006, pp. 785–790.

[29] 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, Jul. 2009.

[30] A. Anjum, S. u. R. Malik, K.-K. R. Choo, A. Khan, A. Haroon, S. Khan, S. U. Khan, N. Ahmad, and B. Raza, "An efficient privacy mechanism for electronic health records," *Computers & Security*, vol. 72, pp. 196–211, Jan. 2018.

hrfoawroauweorui