n-Dependency: dependency diversity in anatomised microdata tables

ANDERS H. LANDBERG, J. WENNY RAHAYU
and ERIC PARDEDE

Department of Computer Science & Computer Engineering La Trobe University, Melbourne, Australia

Abstract

k-Anonymity and l-Diversity have laid the fundamental techniques for preserving privacy in microdata, and many research works have been inspired by them, proposing better and stronger levels of privacy. A common technique for achieving higher privacy in microdata tables is to diversify the records in such a way that sensitive information stored in the data is less likely to be disclosed. While most of the approaches succeed in protecting the original sensitive information to a high degree, issues arise when sensitive values are generalised along a hierarchical taxonomy, causing an increase in probability of privacy disclosure already after the first level of generalisation. This paper introduces n-Dependency, a novel technique that considers the hierarchical nature of sensitive information and their generalisations when diversifying the microdata. We propose a formal model and algorithms, and verify our technique by conducting extensive experiments.

Keywords: data privacy, algorithms, taxonomy, semantic distance

1 Introduction

Privacy preservation in microdata is an important issue when making the data available to the public, for research and general interest purposes. Many privacy preserving properties and methods have been proposed to prevent an adversary from re-identifying a person in a microdata table. An approach called 'anatomy' has introduced a protective method in which quasi-identifying attributes and the sensitive value are separated, producing a k-anonymous and l-diverse quasi-identifier table (QIT). This is achieved without generalising the data, as sensitive values with their respective frequencies are stored in the sensitive table (ST). The advantage of anatomy is that it satisfies the existing properties of k-anonymity and l-diversity without sacrificing loss of precision when aggregating the data in queries.

Example 1: An adversary knows Bob, 23, who lives in a suburb with postcode 11000. The adversary now wants to find out what disease Bob has by looking up the microdata in Table 1. Note that the 'name' attribute (in Tables 1 and 2) are not published and the adversary can not view them. However, having precise information about Bob's profile, the adversary is able to uniquely identify which data record must be Bob's. The adversary can now see that Bob has got pneumonia because he is the only person in that postcode.

Example 2: Again, the adversary has got the same information as in the previous example, but is now restricted to the QIT and the ST tables as shown in Tables 2 and 3. Although the adversary finds Bob's record in QI group 1, they must choose between 4 different possible

E-mail: ahlandberg@gmail.com (Anders H. Landberg), w.rahayu@latrobe.edu.au (Wenny Rahayu), e.pardede@latrobe.edu.au (Eric Pardede)

doi:10.1093/jigpal/jzq015 Advance Access published 10 May 2010

TABLE 1. The microdata

QI#	Name	Age	Gender	ZIP	Disease	ICD-10
1	Bob	23	M	11000	Pneumonia	J12-J18
1	Tom	27	M	13000	Dyspepsia	K30
1	Andy	35	M	59000	Gastritis	K29.[0-7]
1	David	59	M	12000	Bronchitis	J20
2	Alice	61	F	54000	Chronic viral hepatitis	B18
2	Helen	65	F	25000	Mitral stenosis	I05.0
2	Jane	65	F	25000	Multiple valve diseases	I08
2	Lisa	70	F	30000	Acute hepatitis A	B15

TABLE 2. The quasi-identifier table (QIT)

QI#	Name	Age	Gender	ZIP
1	Bob	23	M	11000
1	Tom	27	M	13000
1	Andy	35	M	59000
1	David	59	M	12000
2	Alice	61	F	54000
2	Helen	65	F	25000
2	Jane	65	F	25000
2	Lisa	70	F	30000

TABLE 3. The sensitive table (ST)

QI#	Disease	Count	ICD-10
1	Pneumonia	1	J12-J18
1	Dyspepsia	1	K30
1	Gastritis	1	K29.0-K29.7
1	Bronchitis	1	J20
2	Chronic viral hepatitis	1	B18
2	Mitral stenosis	1	I05.0
2	Multiple valve diseases	1	I08
2	Acute hepatitis A	1	B15

sensitive values for this group. These are pneumonia, dyspepsia, gastritis, and bronchitis. This means that Bob could have either one of these diseases, with the even likelihood of 25%. Note that QIT is 4-anonymous (i.e. it contains 2 groups a 4 records), 4-diverse, and 4-invariant (i.e. all records in each group have distinct sensitive values).

The values in the ''ICD-10'' - column in Table 1 represent real disease codes as found in the ICD-10 taxonomy [1].

1.1 Microdata tables

In the literature, un-aggregated statistical person data is commonly referred to as microdata. For example, the bureau of statistics publishes sets of census (micro)data that can be publicly

[1] http://www.who.int/classifications/icd/en/

downloaded and used for analysis and research. By analysing definitions for microdata in the literature, we identify the following attributes that are common to microdata sets that store information about persons.

- Each person in the data has a unique identifier (e.g. SSN) which is removed when the data is published
- Each data record has a set of quasi-identifying (QI) attributes that can be used to re-identify the person (e.g. dob, gender, zipcode)
- Each data record has a set of sensitive attributes that must not be linked with a particular person (e.g. health condition)

Unique identifiers are keys such as social security number, driver's licence number, bank account number. Using such a key will lead to successful identification of individuals in the microdata, and thus disclose the sensitive attributes of these individuals. An adversary can use the unique identifier to link several datasets on which it exists, and in this way filtering out a list of individuals that are common on all datasets. This type of privacy attack is referred to as *linking attack* and will be discussed in more detail later in the paper. To overcome this obvious privacy flaw, unique identifiers are removed from the data before it is published.

After removing unique identifiers, an individual cannot be re-identified by a single attribute in the data. However, by combining other attributes that are much easier to obtain than unique keys, the sets of individuals that share common attributes can be narrowed down and be used in linking attacks. In the case where only one individual remains in a set, this individual is successfully re-identified. For this reason, any attributes that are not unique keys are referred to as QI-attributes. Also sensitive attributes can be used as QI-attributes.

Sensitive attributes store information about individuals that should not be disclosed. Consider the following example. Alice is an HR manager and Bob is applying for a job in her company. From Bob's CV Alice obtains a set of QI-attributes such as date of birth and zip code. Also knowing that Bob is male, she accesses a publicly available health database published by the state government and queries it with Bob's attributes. Ten records are returned by the query that have common QI-attributes. Alice uses her background knowledge to narrow down her search. From Bob's CV she also knows that he has a tertiary degree. Using this information combined with the other QI-attributes on a census dataset, she finds that there is only one Bob with these attributes who has a tertiary degree. Thus, it must be Bob's record. From this dataset, Alice learns about two other QI-attributes related to Bob, which she can use to query the health database again. After a new query against the health database with five QI-attributes, the search results in one record being returned. Alice has successfully re-identified Bob's record in the health database and disclosed his sensitive attribute (health condition). Seeing that Bob suffers from a heart problem, she decides not to hire him.

This section has given an introduction to the fundamentals of privacy preserving models, and mentioned some existing issues. Next, we focus on some of the outstanding defects of current approaches and explain our motivation on how to overcome the problems.

1.2 Defects of l-diverse and m-invariant tables

Although anatomy presents an ideal concept of preventing disclosure of sensitive information, yet maintaining accuracy in aggregate analysis, it does not consider the fact that the sensitive

TABLE 4. The 2-dependent quasi-identifier table (QIT)

QI#	Name	Age	Gender	ZIP
1	Bob	23	M	11000
1	Tom	27	M	13000
1	Alice	61	F	54000
1	Helen	65	F	25000
2	Andy	35	M	59000
2	David	59	M	12000
2	Jane	65	F	25000
2	Lisa	70	F	30000

values may be related, or dependent on the same parent, and as such, reveal more information about possible sensitive values. This issue can not be avoided by k-anonymous, l-diverse, m-invariant, t-close, nor anatomised tables, as it is subject to extended knowledge about the sensitive values. In the case of inpatient health data, this extended knowledge is represented by ICD-10 codes available at the International Statistical Classification of Diseases and Related Health Problems [2], which is publicly available information.

The following example shall explain this issue in detail. From looking at tables 2 and 3, and consulting the ICD-10 codes, the adversary finds that pneumonia and bronchitis are diseases of the respiratory system. Further, it is known that dyspepsia and gastritis are diseases of the digestive system. Although this knowledge about the sensitive values in QI-group 1 do not help to specify which exact disease Bob has got, it helps to deduct the following information.

Bob has either got a disease with his digestive system, or with his respiratory system, with the even likelihood of 50%. This could pose a problem, because often it is sufficient to know 'roughly' what disease a person has, i.e. the next approximate category, in order to make further conclusions or decisions in response to the newly attained knowledge.

Health information is a good example for this issue, because very closely related diseases can have different names, and as such are not identified as non-diverse. An example of this could be 'blighted ovum' and 'missed abortion', which lexically look completely different, however, both belong to the disease group 'pregnancy with abortive outcome'.

1.3 Motivation

To overcome the defects of l-diverse and m-invariant anatomised tables, we introduce a novel technique *n-dependency*, which produce QIT and ST tables that captures dependency amongst sensitive values and achieves a higher level of diversification.

In Tables 4 and 5, data records 3,4 have been swapped with data records 5,6, hence further diversifying the sensitive values in both QI-groups. From Table 5, we can see that each QI group contains sensitive values that belong to different disease groups. For the sake of illustration, we have included the ICD-10 codes in the ST tables. If the adversary was to make a choice of Bob's disease now, they would have to choose between pneumonia, dyspepsia, chronic viral hepatitis, and mitral stenosis.

After generalising the sensitive disease values one step higher, the choice would be between respiratory disease, digestive disease, viral hepatitis, and chronic rheumatic heart diseases.

[2]http://www.who.int/classifications/icd

TABLE 5. The 2-dependent sensitive table (ST)

QI#	Disease	Count	ICD-10
1	Pneumonia	1	J12-J18
1	Dyspepsia	1	K30
1	Chronic viral hepatitis	1	B18
1	Mitral stenosis	1	I05.0
2	Gastritis	1	K29.0-K29.7
2	Bronchitis	1	J20
2	Multiple valve diseases	1	I08
2	Acute hepatitis A	1	B15

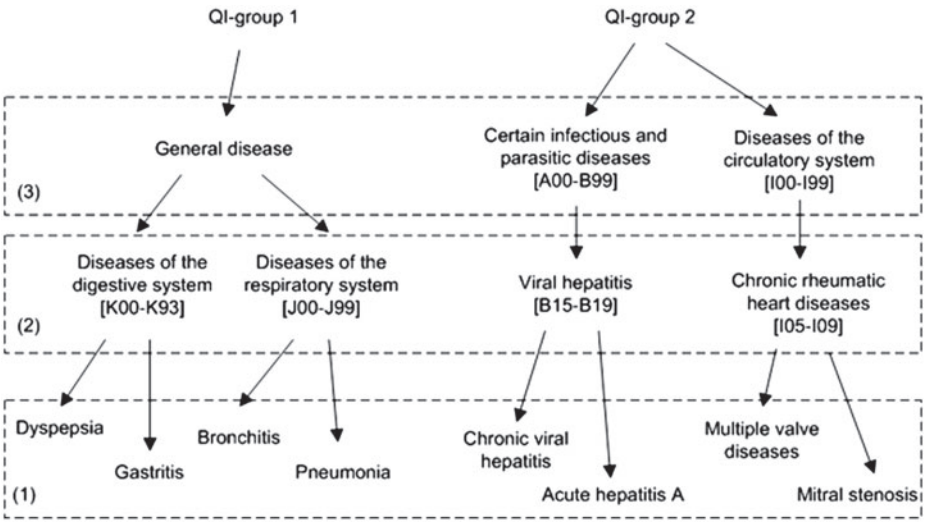

FIG. 1. 4-diverse, 1-dependent QI-groups

These disease categories are clearly more diverse, and all other privacy preserving properties are still fulfilled. In fact, the nearest common disease group of these four values would be 'disease', which comprises all other subgroups.

This means that for QI-group 1, the sensitive values can not be diversified any further, as all dependencies have been removed. The resulting QIT table is said to be 2-dependent, as the distance from each sensitive value to any other sensitive value in that group is at least two levels (i.e. two disease groups) away. For this reason, an additional level of generalisation of the sensitive values does not change their diversity.

1.4 Rationale of n-dependency

The goal of n-dependency in a microdata table is to remove parent-child dependencies and subsequently ancestor-descendant dependencies of sensitive values. To achieve this, let us look at Figures 1 and 2, where the process of removing dependencies is illustrated.

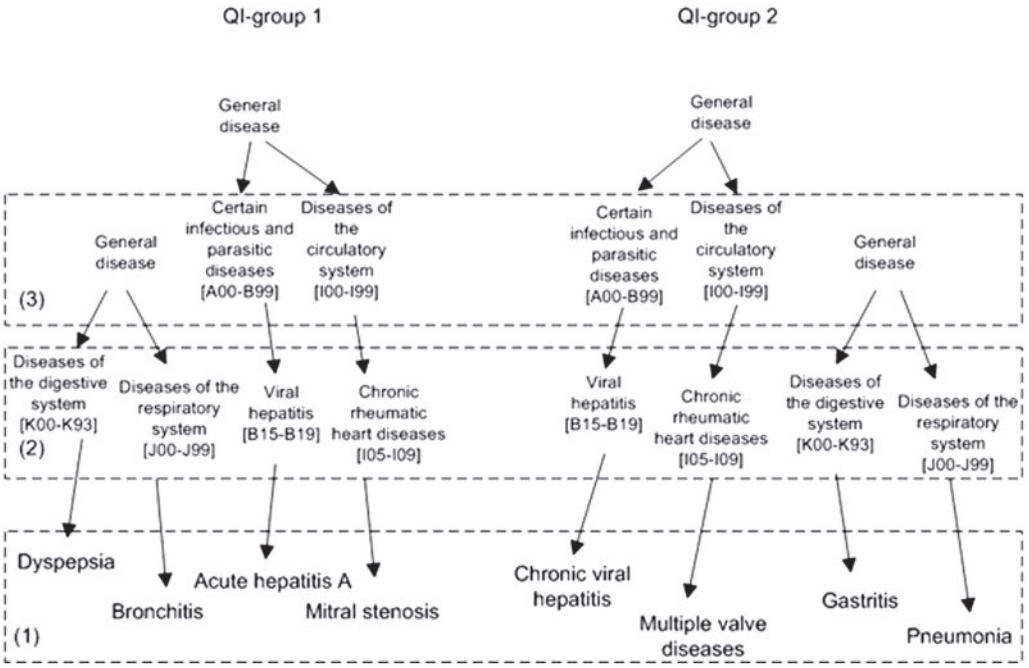

FIG. 2. 4-diverse, 2-dependent QI-groups

In Figure 1, the sensitive values as stored in the microdata are represented by leaf nodes (1), the respective parents of each disease is represented in (2), and the grandparent in (3). As clearly can be seen, both QI-groups are 4-diverse, in fact they are 4-invariant, as all sensitive values are distinct. However, when looking at the next higher level of generalisation (2), then each QI-group is only 2-diverse, because two sensitive values have a common parent in each group. This means that an adversary will be able to narrow down the possible next common disease groups by 2, and hence have a 50% chance to determine the next common disease group for a patient.

To overcome this shortcoming, we can start to remove dependencies in QI-group 1. First, we identify a leaf node in QI-group 2, which does not have the same parent as dyspepsia and gastritis (from (1)). Then, we swap these nodes together with their ancestors (i.e. we swap branches) to increase the diversity of the second level of generalisation in QI-group 1. By following these steps, we finally end up with Figure 2.

In Figure 2, all parent-child dependencies have been removed in respect to the leaf nodes, such that in each QI-group, none of the leaf nodes have the common parent. As a result of this, the first level of generalisation (2) is now 4-diverse, and it takes 2 levels of generalisation (3) to find a common parent for any of the leaf nodes in (1).

A further attempt to diversify the second level of generalisation (3) proves to be impossible, as dyspepsia and gastritis have a common ancestor in (3), and therefore this level of generalisation is only 3-diverse. We can therefore deduct that a QI-group is n-dependent, if it's (n-1)th level of generalisation is n-diverse, or, in other words, if the closest common ancestor for any of its leaf nodes is at least on the n-th level of generalisation.

1.5 Contributions

This paper presents a systematic study of the n-dependency technique. First, we formalise our model based on the l-diversity and anatomy techniques. We show that each pair of QIT and ST in an n-dependent table remains distinct in its QI-group, when the sensitive values in that QI-group are at least n levels of generalisation distant from each other.

Second, we develop an algorithm based on the anatomy algorithm that constructs anatomised tables QIT and ST and maximises n-dependency amongst the QI-groups.

Finally, we conduct extensive experiments to measure the effectiveness and efficiency on real data sets. The rest of the paper is organised as follows. We first highlight the issue with a motivating example and then give an overview over related work. Section 3 proposes the formal model, Section 4 explains the theoretical foundation of our approach, and Section 5 presents our experiments. Finally, Section 6 concludes the paper and gives foresight for future work.

2 Background

2.1 Related Work

k-anonymity. By definition, a set of data is said to be k-anonymous when each QI-group contains at least k tuples [24]. QI stands for Quasi Identifier, and a QI-group is the set of attributes that can be used to identify a sensitive value within the tuple. With increasing k, the ambiguity of the data will increase, but the detail of the data will decrease, as specific values are generalised into ranges. The procedure commonly used to achieve k-anonymity is generalisation, which specifies ranges of values for attributes in such a way, that in each QI-group there are at least k tuples, and hence, the data set is k-anonymous.

k-anonymity [5] [3] [7] [34] [8] [28] [17] [27] [9] [15] [4] [10] [1] [19] [36][22] [21] [14] [33] [11] [12] [26] [25] [2] [32] [30] is a popular approach to data privacy and many works have been proposed in this area. It uses generalisation (and suppression) to de-identify the data. In this process, the values of the QI attributes are generalised into ranges that individual QI combinations cannot be associated with one particular sensitive value. k-anonymity ensures that in each QI-group there are at least k tuples that share the common ranges of quasi identifier values.

l-diversity. l-diversity [18] [6] [35] [20] [16] was proposed to overcome the shortcomings of k-anonymity. By definition, a set of data is said to be l-diverse if in each QI-group, at most 1/l of the tuples possess the most frequent sensitive value. As a consequence, there are at least l distinct sensitive values in each QI-group in an l-diverse data set. This property ensures that a set of quasi-identifiers cannot be deterministic for a particular sensitive value, which would be the case if for a QI-group, l=1. It ensures the diversity of sensitive values within each QI-group, which is not previously addressed by k-anonymity.

l-diversity uses the diversity of sensitive values within QI-groups to strengthen the degree of privacy protection. It ensures that in each QI-group the most frequent sensitive value appears in at most 1/l of the tuples. This also implies that each QI-group must have at least l distinct sensitive values.

Anatomy. Xiao et al. propose to separate the QI-groups from the sensitive values, by linking them together with group-ids [29]. In this way, the data values in the QI-groups don't need to be generalised, hence data detail is preserved, and k-anonymity and l-diversity

are still ensured. This approach utilises two external lookup-tables that hold (i) the quasi-identifier group tuples and a group-id, and (ii) the sensitive values, a count for the frequency of each sensitive value, and a group-id. While anatomy does not directly improve the degree of privacy for the data, it ensures higher detail (granularity) of the data in conjunction with existing privacy properties.

Anatomy uses external lookup tables to separate the data from the sensitive values. This technique is applied for k-anonymous and l-diverse data sets.

t-closeness. Li et al. propose further improvements to the existing privacy properties k-anonymity and l-diversity [13] [23]. The approach is based on the distribution of each sensitive value in the QI-groups versus their global distribution in the data set. By definition, a QI-group is said to have t-closeness if the distance of the distribution of each sensitive value in this group and the global distribution of the sensitive value is no more than threshold t. Further, a data set is said to have t-closeness if all its QI-groups have t-closeness. The distance of distributions is calculated using the Earth-Mover's distance (EMD), which is a Monge-Kantorovich transportation distance [5] in disguise, and calculates the difference in work that is necessary to transform one distribution to another. For example, t-closeness ensures that a particular sensitive value cannot occur in only one QI-group (when the data set has got, say 10 QI-groups), because the distributions of the sensitive value in the entire data set and the QI-group where it is contained, are too distant, given a reasonable distance threshold t.

t-closeness uses the Earth Mover's distance to measure whether sensitive values in the entire data set and each QI-group are within a threshold t. This approach can be used to measure the degree of privacy protection after k-anonymity and l-diversity.

m-invariance. Xiao et al. propose a method that ensures privacy in dynamic aggregated microdata, after insert and delete operations on the data have been performed [31]. This approach overcomes the re-publication issues that occur with k-anonymous and l-diverse data sets. A re-publication issue appears when differences in two published data sets (both being k-anonymous and l-diverse) lead to identifying information for a particular tuple.

For example, a QI-group contains two tuples with sensitive values 'heart disease' and 'lung cancer', respectively. In a later release of the data, the second tuple has been deleted, and another tuple has taken its place in the same QI-group. Now, the sensitive values are 'heart disease' (first tuple) and 'bronchitis'. If an adversary knows in advance that a person in this QI-group has got either heart disease or liver disease, then the information that is gained from comparing data publications before and after, is identifying for the first tuple. m-invariance addresses this issue by introducing counterfeit values, which act as dummy-values to substitute deleted tuples, in the case of critical absence of tuples.

By definition, a data set is said to be m-invariant, if throughout multiple republications, each QI-group contains at least m tuples, and in each tuple there are m distinct sensitive values. Therefore, m-invariance implies m-diversity, but not vice versa.

m-invariance is a modification of l-diversity in the sense that it ensures all sensitive values in each QI-group to be distinct. This property is enforced using counterfeit tuples after insertions and deletions of tuples in the data set.

This section has given an overview over related work in the area of privacy preservation in microdata. Next, we will discuss families of the most common privacy attacks and how existing methods approach them.

2.2 Privacy Attacks

Linking attack. Initial attempts to protect published microdata against disclosure of individuals' sensitive values were to remove explicit identifiers such as social names, phone numbers and addresses. However, the remaining attributes such as age, gender and zip code can be combined, and in a majority of cases again lead to privacy disclosure where a unique combination of these values is correlated with a particular sensitive value. Using these quasi identifying values and linking them to external datasets such as voting lists, individual's identities (found on the voting list) can be uniquely matched against the sensitive value (found in the microdata). In response to this problem, k-anonymity was proposed which aims at removing the threats of linking attacks.

Homogeneity attack. This is also known as similarity attack. A homogeneity attack exploits poor diversity of sensitive values within a QI group, i.e. when a high percentage of records in a QI group share the same sensitive value. Consider the often used example where Alice wants to find out Bob's medical condition. Alice and Bob are neighbours, so Alice knows Bob's age, gender and zip code. Querying the microdata and using her knowledge of Bob's quasi identifiers as query filters, she finds the QI group in which Bob's record is contained. Because all records in this QI group have the same sensitive value, Alice can determine with 100% certainty, which medical condition. This example shows us that anonymisation alone is not sufficient to enforce strong data privacy.

Background knowledge attack. An adversary may have background knowledge that can help to link a tuple's QI values to a sensitive value. This could be specific information about an individual which will help to guess the right one among several sensitive values, or general information about a group of individuals, such as the fact that women cannot have prostate cancer. Among all the various kinds of attack, this is probably the most difficult one to predict, since background knowledge of potential adversaries is often limited or even unknown. Therefore, it becomes increasingly difficult to model this background knowledge and to use it as weighting factor in privacy models.

Skewness attack. While diversification of sensitive attribute values improves the privacy protection significantly by reducing the possibility of a homogeneity attack, an issue arises when the sensitive values in equivalence classes are distinct, but semantically similar. An adversary can learn important information from semantic similarities of values within QI-groups, because some sensitive values may have stronger semantic relationships than others, and so they can be clustered and invalidate the initial diversity privacy preserving property. Also, an ill-formed distribution of similar sensitive values across QI-groups and their overall distribution in the microdata may cause serious leaks for attacks by skewness.

3 Formal Model

A microdata table T contains d quasi-identifier (QI) attributes A_1^q, A_2^q, ..., A_d^q and a sensitive value A^s. For any tuple $t \in T, t[i] = A_i^q$ (QI values) for $1 \leq i \leq d$ and $t[d+1] = A^s$ (sensitive value). Sensitive values are mapped into a taxonomy tree T_X with a mapping function $f_X(A^s) = TN_{idx}^p$ that assigns every sensitive value to a tree node TN with parent node index p and index idx. The root node of T_X is denoted as TN_0^{-1}. Index -1 specifies no further parents, and index = 0 specifies the root level. Further, we use $L(TN_{idx}^p)$ to denote the tree level of a node. Thus, $L(TN_0^{-1}) = 0$, with incremental levels for child nodes.

TABLE 6. Example QI-groups produced by Anatomy

QI#	A^s	$G(A^s)$	div_0^q	$div_{\lambda_{final}}^q$	2-dep
1	Mul. Valve Dis. Gastritis Viral Hepatitis Dyspepsia	Circulatory Disease Digestive Disease Infectious Parasitic Digestive Disease	4	3	X
2	Mul. Valve Dis. Gastritis Bronchitis Pneumonia	Circulatory Disease Digestive Disease Respiratory Disease Respiratory Disease	4	3	X
3	Viral Hepatitis Chr. Viral Hepat. Mul. Valve Dis. Dyspepsia	Infectious Parasitic Infectious Parasitic Circulatory Disease Digestive Disease	4	3	X
4	Flu Gastritis Bronchitis Pneumonia	Respiratory Disease Digestive Disease Respiratory Disease Respiratory Disease	4	2	X
5	Chr. Rh. Heart Dis. Viral Hepatitis Chr. Viral Hepat. Mul. Valve Dis.	Circulatory Disease Infectious Parasitic Infectious Parasitic Circulatory Disease	4	2	X

DEFINITION 1. **(Generalisation)** A *generalisation* G_k of n sensitive values $A_1^s, ..., A_n^s$ is defined as their *k-th* predecessor in T_X such that $G_k(A_1^s) = ... = G_k(A_n^s)$ given that $L(f_X(A_1^s)) = ... = L(f_X(A_n^s))$.

Example: Given that $f_X(A^s) = TN_{idx}^{p1}$, we obtain the *1st generalisation* of A^s as TN_{p1}^{p2}, where *idx* is the tree level of A^s and *p1* is the tree level of $G_k(A^s)$ and *p2* is the tree level of the generalisation's parent. Therefore, a sensitive value A^s can have at most $L(f_X(A^s))$ generalisations.

DEFINITION 2. **(QI-group)** A QI-group QI is defined as a subset of T, such that $\bigcup_{i=1}^{m} QI_i = T$ where $QI_j \cap QI_k = \oslash$ and $j, k \in \{0..m\}$.

A QI-group is a subset of T. The union of all QI-groups equals T. The intersection of any two QI-groups is the empty set. Each QI-group has at least k tuples, where k specifies the level of anonymity.

DEFINITION 3. **(l-diverse QI-group)** A QI-group is defined as *l-diverse*, if at most $\frac{1}{l}$ of the tuples in the QI-group contains the most frequent sensitive value. In an *optimal l-diverse* QI-group, *all* sensitive values are distinct. This property is also referred to as *m-invariant* [31].

Example: Tables 6 and 7 show how Anatomy and n-Dependency techniques have arranged tuples into five QI-groups. In all groups that are generated by both techniques, the A^s values are distinct. As each group contains exactly 4 tuples, they all satisfy the property of 4-diversity. Hence, the table is said to be 4-diverse because all its QI-groups are 4-diverse.

TABLE 7. Example QI-groups produced by 2-Dependency

QI#	A^s	$G(A^s)$	div_0^q	$div_{\lambda_{final}}^q$	2-dep
1	Dyspepsia Bronchitis Viral Hepatitis Chr. Rh. Heart Dis.	Digestive Disease Respiratory Disease Infectious Parasitic Circulatory Disease	4	4	√
2	Dyspepsia Bronchitis Viral Hepatitis Mul. Valve Dis.	Digestive Disease Respiratory Disease Infectious Parasitic Circulatory Disease	4	4	√
3	Gastritis Pneumonia Viral Hepatitis Mul. Valve Dis.	Digestive Disease Respiratory Disease Infectious Parasitic Circulatory Disease	4	4	√
4	Gastritis Pneumonia Chr. Viral Hepat. Mul. Valve Dis.	Digestive Disease Respiratory Disease Infectious Parasitic Circulatory Disease	4	4	√
5	Gastritis Flu Chr. Viral Hepat. Mul. Valve Dis.	Digestive Disease Respiratory Disease Infectious Parasitic Circulatory Disease	4	4	√

DEFINITION 4. **(Sensitive values diversity)** The *sensitive values diversity* div_k^q of a QI-group q for the *k-th* generalisation of the sensitive attribute is defined as the number of distinct sensitive values $div_k^q = \text{distinct } \left|G_k(A^s)\right| \text{ in } q$.

Example: Referring again to Tables 6 and 7, we observe that all QI-groups generated by both techniques have $div_0^q = 4$. After one level of generalization of A^s, all generalized values of A^s in the n-Dependency generated table remain distinct in each QI-group, while the Anatomy generated table suffers from diversity loss in all QI-groups. In fact, QI-groups 4 and 5 are now merely 2-diverse.

DEFINITION 5. **(n-dependent QI-group)** A QI-group is defined as *n-dependent*, if the closest common generalisation of all sensitive values in the QI-group is found not closer than on generalisation G_n, thus ensuring that A^s and its *n-1* generalisations have equal sensitive values diversity *div*.

Example: In Table 6 all QI-groups are 1-dependent because they do not have the same sensitive values diversity *div* for A^s and $G_1(A^s)$, whereas the n-Dependency generated table (Table 7) is arranged in such a way that $A^s = G_1(A^s)$ for all sensitive values. As such, all QI-groups in Table 7 satisfy 2-dependency.

We would like to note that *n-dependency* can be achieved on both l-diverse tables, as well as m-invariant tables. n-dependency implies that the further levels of generalisation of sensitive values maintain the same level of l-diversity or m-invariance as the original sensitive value, or the previous generalisation thereof.

n-dependency increases the privacy protection of generalisation of sensitive values, not of the original sensitive values themselves.

4 Theory of n-Dependency

Dependency diversity is the key idea behind n-Dependency. The goal is to rearrange tuples into QI-groups in such a way that after the tuples' sensitive values are generalised $n-1$ times, the generalised values remain distinct. This section explains the required methodology for generating n-Dependent tables.

4.1 Pre-processing of taxonomy tree

To achieve distinct generalized values of sensitive attributes, three steps are necessary. First, all sensitive values in the microdata must be mapped into an (existing) taxonomy tree T_X and their frequencies must be counted and attached to their respective tree node TN in T_X (Mapping step). Also, each node that relates to a value in the microdata (has a frequency greater than 0) must be marked. After this step T_X contains three types of nodes: (i) nodes that have been marked as existent in the microdata and having a frequency greater than 0, (ii) nodes that have not been marked but that have children that exist in the microdata, and (iii) nodes that have not been marked and do not have any children that exist in the microdata.

DEFINITION 6. **(Initial diversification level)** The initial diversification level $\lambda_{initial}$ is defined as the tree level $L(TN_i)$ in T_X where $\forall TN_i : f_X(A^s) \neq G_1(TN_i)$(ancestor/generalization does not exist in microdata) and $\exists TN_i : f_X(A^s) = TN_i$, where i is the number of nodes on $\lambda_{initial}$.

Second, the frequencies of all nodes are recursively propagated to their parents, thus aggregating the frequencies along the descendant-ancestor axis in the tree (Pruning step, Figure 3, Algorithms 1, 2). After this step all nodes have been marked with their respective tree level and those that have values existing in the microdata and their ancestor nodes have been pruned with aggregated frequencies. All other nodes remain irrelevant for further steps.

Algorithms 1 and 2 recursively perform a *depth-first traversal* on T_X and aggregate the node frequencies. All nodes whose values are in the microdata are marked. This is an important indicator because it will be required to find the initial diversification level and later the highest degree of n-dependency.

Third, the set of nodes on which to apply the initial dependency- diversification must be determined (Finding initial diversification level, Definition 6, Algorithm 3).

Algorithm 3 iteratively performs a *breadth-first traversal* on T_X and finds the lowest tree level on which the sensitive values can initially be diversified. The initial diversification level is the first tree level on the breadth-first traversal that contains a node whose value is in the microdata.

The tree level $\lambda_{initial}$ for initial diversification will determine which sensitive values must be generalised. Also, it can now be determined whether or not a diversification on $\lambda_{initial}$ can be achieved, which depends on the frequencies of all TN on $\lambda_{initial}$. If the frequencies of nodes on $\lambda_{initial}$ allow the tuples to be rearranged in such a way that the QI-groups satisfy

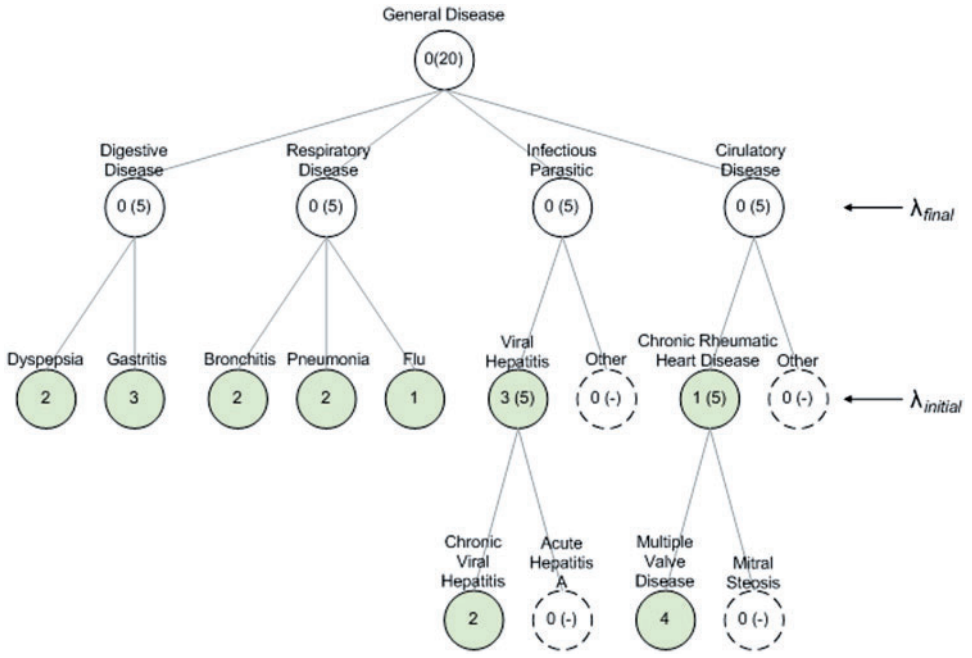

FIG. 3. Pruned T_X for sensitive attribute 'disease'

Algorithm 1: pruneNodesRec

```
Result: Aggregated TN frequency
begin
    currentTreeLevel ← currentTreeLevel + 1
    TN ← Pop stack
    Mark TN as currentTreeLevel
    if TN.frequency > 0 then
        Mark TN as existing in microdata
    if TN is leaf then
        currentTreeLevel ← currentTreeLevel + 1
        return frequency
    foreach TN children do
        Push child on stack
        TN.frequency ← TN.frequency + pruneNodesRec()
    currentTreeLevel ← currentTreeLevel + 1
    return TN.frequency
end
```

l-diversity, then this implies that n-Dependency can possibly be achieved for n > 1, otherwise the Anatomy algorithm will perform equally well as n-Dependency.

After these three steps, the taxonomy tree is ready to be used in the n-Dependency diversification process. Next, we find the maximum value for n, which will correspond to a

Algorithm 2: pruneNodes

```
Data: T_X : Taxonomy tree
Result: Pruned T_X
begin
    currentTreeLevel ← -1
    Empty the stack
    Push T_X root node on stack
    while stack not empty do
        pruneNodesRec()
end
```

Algorithm 3: findLevel

```
Data: T_X : Taxonomy tree
Result: Initial diversification level λ
begin
    λ ← -1
    TN ← T_X root node
    Push TN into queue
    while queue not empty do
        TN ← Remove first from queue
        if TN in microdata then
            return treeLevel of TN
        if !TN.isLeaf then
            Push all TN children into queue
    return λ
end
```

$\lambda_{final} \geq \lambda_{initial}$. Next, all sensitive values that are on nodes below λ_{final} must be generalised to their respective ancestors on λ_{final}. In practice we create an additional column in the microdata table and update it with the generalised values. Finally, we apply the diversification algorithm to the microdata.

Algorithm 4 finds the maximum level of n-Dependency that can be achieved in the microdata given the taxonomy tree T_X and a parameter l that specifies the QI-group size and also the diversity level of the resulting table. According to the algorithm, n-Dependency can be achieved under the condition that the *n-1-th* generalisations of sensitive values on λ_{final} satisfy the *l-diversity* property (Proposition 1).

4.2 Propositions

PROPOSITION 1. **(Achieving n-Dependency)** Given a microdata table T, a taxonomy tree T_X, and a QI-group size l, a dependency level $n = \lambda_{initial} - \lambda_{final}$ can always be achieved if $\lambda_{initial} > \lambda_{final}$.

Algorithm 4: findMaxN

```
Data: T_X : Taxonomy tree
λ : Initial diversification level
l : QI-group size/diversity level
Result: n_max
begin
    n_max ← 1
    sumFreq ← Sum of all node's frequencies on λ
    while λ ≥ 0 do
        foreach TN in T_X on λ do
            ratio ← TN.frequency / sumFreq
            if ratio > 1/l then
                return n_max
        n_max ← n_max + 1
        λ ← λ - 1
    return n_max
end
```

PROOF. Both diversification levels $\lambda_{initial}$ and λ_{final} ensure that generated QI-groups have equal sensitive values diversity (Definitions 4,6). Therefore, n generalizations of the values on $\lambda_{initial}$ are necessary to obtain the values on λ_{final} (Definitions 5,6). Thus, the values on $\lambda_{initial}$ satisfy the conditions to generate n-dependent QI-groups and for the table T to satisfy the n-dependency property ∎

PROPOSITION 2. **(Random assignment)** n-Dependency will always generate a number of n-dependent QI-groups larger than or equal to the number of n-dependent QI-groups generated by approaches that are based on diversification, such as l-diversity and anatomy. The maximum achievable level of dependency in table T be n_{max} so that $p_{n_{max}} = 100\%$ using n-Dependency, then the percentage of QI-groups generated by an alternative approach satisfying $n_{max}..n_1$-dependency are $q_{n_{max}}..q_{n_1}$ where $q_{n_{max}}+..+q_{n_1} = 100\%$ and $p_{n_{max}} \geq q_{n_{max}}$.

PROOF. n-Dependency diversifies the data records based on their maximal generalized values, thus maximizing the semantic distance between the records in each QI-group. Diversification based approaches like l-diversity and anatomy diversify data records based on the initial sensitive value, thus generating QI-groups with random semantic distance. In the best case, this random assignment yields the same effectiveness as n-dependency ∎

4.3 Example

An n-dependent QI-group q maintains its level l of diversity throughout *n-1* levels of generalisation of the sensitive value, such that l is constant for A_j^s, where $1 \leq j < n$.

Under the generalisation error e that is inflicted by the generalisation, the probability p to identify a particular sensitive value generalisation A_i^s is equal to the probability of identifying a particular sensitive value of any previous generalisation, or the original sensitive value itself.

TABLE 8. No n-dependency

$G_1(A^s)$	Respiratory disease		Digestive disease	
A^s	Pneumonia	Dyspepsia	Gastritis	Bronchitis
Query 1	25%	25%	25%	25%
Query 2	50%		50%	

TABLE 9. 2-dependency

$G_{A^s}(A^s)$	Respiratory dis.	Digestive dis.	Viral hepatitis	Heart disease
A^s	Pneumonia	Dyspepsia	Hepatitis A	Valve disease
Query 1	25%	25%	25%	25%
Query 2	25%	25%	25%	25%

Let us assume that Table 3 stores generalisations of the sensitive attribute 'Disease'. We then join Tables 2 and 3 by the group id, and perform the following queries:

// select the sensitive value
(1) SELECT A^s FROM T WHERE $pred(A_1^q)$ AND ... AND $pred(A_1^q)$

// select the first generalisation
(2) SELECT $G_1(A^s)$ FROM T WHERE $pred(A_1^q)$ AND ... AND $pred(A_1^q)$

The query results can be displayed on a two dimensional matrix:

From Table 8, we notice a loss of privacy correlation, as the query result values have diminished to 2 from previously 4. Thus, generalising A_0^q by one level reduces the diversity of the sensitive attribute by 50% and therefore increases the probability to disclose the sensitive attribute from 25% to 50%. In comparison, we perform both queries on the resulting join between Tables 4 and 5, and again display the results in a matrix.

In this scenario (Table 9), privacy correlation is maintained, as the probability of re-identifying the disease of any patient is still 25%. In addition to that, data correlation is not diminished, as there are 4 possible diseases for the patient, and hence the precision remains the same as in the first query.

We define the loss of diversity between the diversity of the original sensitive value and a generalised value on the *n-th* level in QI-group QI_i as the

$$\textit{relative diversity error } err_{div}^n = \frac{\left|div_0^{QI_i} - div_n^{QI_i}\right|}{div_0^{QI_i}}$$

where $1 \leq i \leq$ number of QI-groups.

If $err_{div}^n = 0$ can be achieved for all QI-groups for a dependency level n, then the table is said to be *optimal n-dependent*.

5 Experiments

In our experiments we use a dataset CENSUS[3] that holds US census data from the year 2000 with the attributes shown in Table 10. The dataset has a cardinality of 100k records, containing personal information about 100k American adults. Parameters and tested values can be found in Table 11.

[3]Downloadable at https://international.ipums.org/international/

TABLE 10. Summary of attributes

Attribute	Distinct values	Type
Age	91	quasi-identifier
Gender	2	quasi-identifier
Maritial	6	quasi-identifier
Birthplace	100	quasi-identifier
Employment status	6	quasi-identifier
Class of worker	9	quasi-identifier
Education	15	sensitive
Total income	50	sensitive

TABLE 11. Parameters and tested values

Parameter	Values
dependency n	2, 3, 4
l	4, 5, 6, 7, 8, 9
cardinality	100k
number of QI attributes d	6
query dimensionality qd	2, 3, 4, 5, 6
expected selectivity s	5%

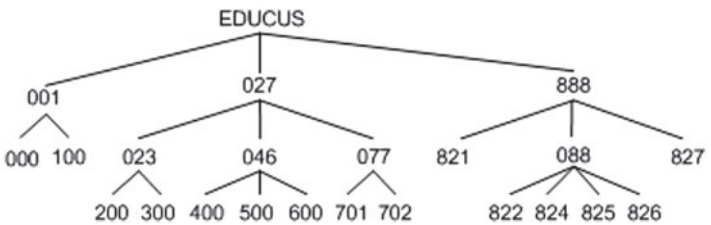

FIG. 4. Taxonomy tree for sensitive atribute 'educus' in EDU-6

From CENSUS, we create two sets of microdata tables, named EDU-6 and INC-6 respectively, where 6 denotes the number of QI attributes. These tables have *Education* (*Income*) as their sensitive attribute, whose values are mapped into taxonomy trees (taxonomy for EDU shown in Figure 4). INC-6 sensitive values have been divided into even intervals and then grouped by 2. Both tables are equal in cardinality (100k) and QI-attributes. The number codes in the taxonomy trees correspond to the category codes as found in the original IPUMIS dataset.

We compare n-dependency against anatomy, which is an approach that dissects diversified tables into quasi-identifier and sensitive tables. We believe that anatomy is among the most convincing approaches to privacy preservation. The performance graphs that follow denote our approach as n-dependency, and the anatomy approach as l-diversity. The reason for this is because anatomy is in effect a version of l-diversity (as is n-dependency).

The first goal of the experiment is to show that existing diversification models do not output tables that are diverse enough when their sensitive attributes can be mapped into hierarchical relationships. The graphs in Figure 5 show the percentage of formed n-dependent QI-groups as a function of diversity level l. We observe that EDU-6 (d) is not diverse enough to satisfy complete 2-dependency for $l > 4$, but achieves a maximum of possible 2-dependent

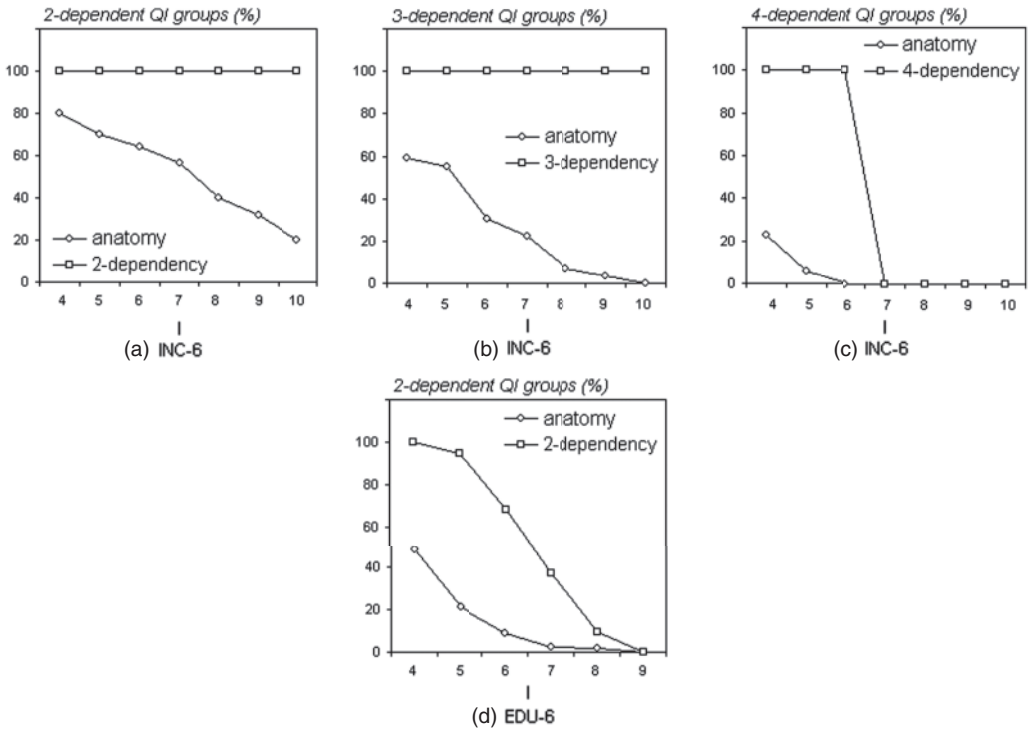

FIG. 5. Dependence Diversity vs. l-Diversity

QI-groups. This measure also determines on how many QI-groups in the table the sensitive value can be inferred with less than 25%, which is 100 - % n-dependent QI-groups. A function value of 0 at l indicates that there are less than l distinct tuples with distinct values on the first generalisation. In INC-6 all tuples have 3 levels of distinct generalisations for $l \leq 10$ (a,b) and $l \leq 6$ (c), which means that generalizing three levels will not increase the probability of inferring the sensitive value.

In the following experiments, the parameter l (applying to both anatomy and n-dependency) is set to 4, implying that the sensitive value of an individual can be correctly inferred with at most 25% probability. For n-dependency, an increasing n will cause this probability to remain stable (in the ideal case constant), thus providing the same protection even after the sensitive values are generalized.

Table 12 lists the code values that correspond to the distinct sensitive values of attribute 'educus' in table EDU-6. Column 'CENSUS' indicates whether the values are actually existent in the CENSUS data table. The values with a 'X' are groupings/generalizations of the data values and are used to build an artificial taxonomy structure as illustrated in Figure 4.

5.1 Aggregate Reasoning

We auto-generate query workloads in bulks of 1000 that have the form:

`SELECT COUNT(*) FROM` Microdata
`WHERE` $pred(A_1^q)$ `AND ... AND` $pred(A_{qd}^q)$ `AND` $pred(A^s)$

TABLE 12. Code-table for EDU-6 taxonomy (educus)

Code	Description	CENSUS
000	NIU (not in universe)	√
100	None or preschool	√
200	Grades 1 to 4	√
300	Grades 5 to 8	√
400	Grade 9	√
500	Grade 10	√
600	Grade 11	√
701	Grade 12, no diploma	√
702	High school graduate or equivalency degree	√
821	Some college, no degree	√
822	Associate degree, occupational	√
824	Bachelors degree	√
825	Masters degree	√
826	Professional degree	√
827	Doctorate degree	√
023	Secondary Junior	X
046	Secondary Senior	X
077	Secondary Grade 12	X
088	Tertiary Degree	X
001	No School	X
027	Secondary	X
888	Tertiary	X

A_1^q, ..., A_{qd}^q are qd random QI-attributes, and $pred(A^s)$ is the sensitive attribute, where qd is the *query dimensionality*. For any attribute A, the predicate $pred(A)$ has the form

$(A = x_1$ OR $A = x_1$ OR ... OR $A = x_b)$

where x_i is a random value in the domain of A (all attributes are discrete). The value of b depends on the *expected query selectivity* s:

$$b = \lceil |A| \cdot s^{1/(qd+1)} \rceil$$

where $|A|$ is the domain size of A. In our experiments we use $s = 5\%$.

Given the CENSUS microdata relation, we compute two sets of anatomized tables that satisfy the anatomy property, and the n-dependency property for $n = 2$ (EDU-6) $2 \leq n \leq 4$ (INC-6). Then, we execute a workload of 1000 queries for each value of $2 \leq qd \leq 6$ and each table and measure the effectiveness of anatomy vs. n-dependency by recording the *average relative query error* and our own proposed metric *average relative diversity error*. The relative query error equals $\frac{|act-est|}{act}$ where act is the actual query result from the microdata, and est the estimated result from the anatomized tables. The relative diversity error equals $\frac{|div_0 - div_n|}{div_0}$ where div_0 is the number of distinct original sensitive values in the QI-group of a result tuple, and div_n the number of distinct sensitive values in the same QI-group after n generalisations.

The first set of experiments shown in Figure 6 measures the query accuracy as a function of qd for both EDU-6 (a) and INC-6 (c). As previously mentioned, we have computed

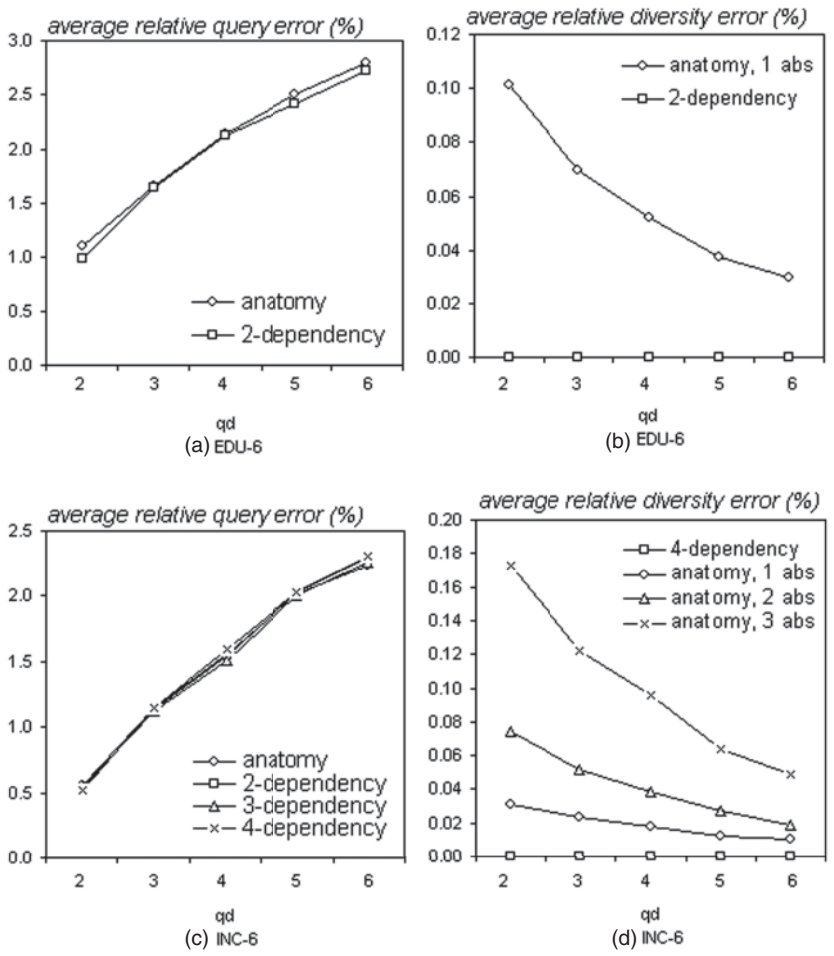

FIG. 6. Query Accuracy/Diversity Accuracy vs. Query Dimensionality

anatomized and (2,3,4)-dependent versions of table INC-6, and executed query workloads on all of them. The graphs in (b) and (d) show that there is almost no difference in average relative query error, thus proving that our technique does not affect the accuracy of aggregate analysis.

Next, we analyse the degree of diversity in query results for both anatomy and n-dependency. For example, a query returns 10 tuples whose A_0^s values are distinct ($div_0 = 10$), its A_1^s values have 5 distinct values ($div_1 = 5$), and its A_2^s values are all equal ($div_2 = 1$), then the diversity error err_{div}^n for the first generalisation is calculated as $err_{div}^1 = \frac{|10-5|}{10} = 0.5$, and for the second generalisation $err_{div}^2 = \frac{|10-1|}{10} = 0.9$.

The highest possible diversity error is calculated by $\frac{|res_{qry}|-1}{|res_{qry}|}$, where $|res_{qry}|$ is the size of the query result set. Further, a high diversity error err_{div}^n directly indicates the probability of disclosure of the sensitive value after n generalisations. In the above example two generalisations are required to determine the generalization of the sensitive value. Because all

generalised sensitive values are equal, they are being disclosed with 100% probability, and thus the privacy attack succeeds.

Returning to the graphs (b) and (d) in Figure 6, we observe that n-dependent tables completely protect against this attack. The maximum levels of n-dependency are 2 (EDU-6) and 4 (INC-6), hence preserving a constant breach probability throughout 1 (3) levels of generalisation for the sensitive values. The anatomy technique performs particularly weak for low *qd*, and with increasing levels of generalisation. This phenomenon is caused by decreasing diversity among sensitive values after increasing number of generalisations, and a random diversification strategy of generalised values by the anatomy technique, which leads to an increase in breach probability.

5.2 Protection against privacy attacks

As our approach is based on the principle of l-diversity, n-dependent tables satisfy the same degree of protection against privacy attacks as described in Section 2.2. However, l-diverse tables (anatomised and/or m-invariant) still do not guarantee that the same level of protection against privacy disclosure is maintained throughout generalised sensitive attributes, and can cause a serious threat.

Especially in deep taxonomies where the information loss between a sensitive value and its next generalisation/s is very small, privacy information can easily leak when QI-groups do not satisfy the n-dependency property. Thus, by separating diversity dependencies we minimise the probability of disclosing sensitive information after generalisation of sensitive attributes. As a result thereof, n-dependency performs particularly well in protecting against homogeneity attacks, background knowledge attacks, and skewness attacks.

Linking attack. Our proposed solution includes removing all unique identifiers from the microdata, thus making it resistant to direct linkage with external datasets. Therefore, by removing unique identifiers we protect the data against linking attacks while preserving the informational value (attributes) of the data.

Homogeneity attack. Poor diversity of sensitive values inflicts risk of successful homogeneity attacks. This also yields for poor diversity of generalised sensitive values. As opposed to l-diversity, anatomy and similar approaches, applying our methodologies guarantees resilience against homogeneity attacks over multiple generalisations of sensitive values. This is so because our methodologies diversify the microdata on the most general possible level, thus minimising the risk of successful homogeneity attacks.

Skewness attack. By separating tuples with semantically similar sensitive values in accordance with the taxonomy tree, we protect the data against skewness attacks. Effectively, this means that all sensitive values in each QI-group will have a maximum distance from one another. As a consequence of this, queries that match these QI-groups will never return result sets with more similar sensitive values than $\frac{1}{l} \times$ *retrieved QI-groups* (Experiments 6 b and d).

6 Conclusion and Future Work

We have shown that diversity based approaches like l-diversity and anatomy have limitations with respect to the diversity of sensitive values in QI-groups when generalisations are used

to infer data disclosure. Our proposed solution approaches this problem by considering the hierarchical order of sensitive values and their frequency distribution in the microdata when diversifying the tuples and finally anatomising them into quasi-identifier and sensitive tables.

This paper has introduced a formal model to express the concepts and a set of theorems and algorithms to prove our methodologies. Further, several sets of experiments on real US Census data show the effectiveness of our approach and point out limitations of related works algorithms.

Future work includes further exploration of the n-Dependency technique, particularly how to deal with tables that only satisfy a very low level of n-Dependency. In such cases, different partitions of the table may satisfy different levels of dependency. Also, we will investigate on different types of taxonomies that apply to the table data, and how varying taxonomy structures influence our technique.

References

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

[2] C. Bettini, X. S. Wang, and S. Jajodia. How anonymous is k-anonymous? look at your quasi-id. In *SDM '08: Proceedings of the 5th VLDB workshop on Secure Data Management*, pages 1–15, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] J. Domingo-Ferrer and V. Torra. A critique of k-anonymity and some of its enhancements. In *ARES*, pages 990–993, 2008.

[4] T. Iwuchukwu and J. F. Naughton. K-anonymization as spatial indexing: Toward scalable and incremental anonymization. In *VLDB*, pages 746–757, 2007.

[5] H. Jian-min, Y. Hui-qun, Y. Juan, and C. Ting-ting. A complete (alpha, k)-anonymity model for sensitive values individuation preservation. In *ISECS*, pages 318–323, 2008.

[6] H. Jian-min, C. Ting-ting, and Y. Hui-qun. An improved v-mdav algorithm for l-diversity. In *ISIP*, pages 733–739, 2008.

[7] W. Jiang and C. Clifton. A secure distributed framework for achieving -anonymity. *VLDB J.*, 15(4):316–333, 2006.

[8] R. J. B. Jr. and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, pages 217–228, 2005.

[9] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD Conference*, pages 49–60, 2005.

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, page 25, 2006.

[11] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *KDD*, pages 277–286, 2006.

[12] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM Trans. Database Syst.*, 33(3), 2008.

[13] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, pages 106–115, 2007.

[14] T. Li and N. Li. Towards optimal k-anonymization. *Data Knowl. Eng.*, 65(1):22–39, 2008.

[15] T. Li, C. Tang, J. Wu, Q. Luo, S. Li, X. Lin, and J. Zuo. k-anonymity via clustering domain knowledge for privacy preservation. 4:697–701, 2008.

[16] Z. Li and X. Ye. Privacy protection on multiple sensitive attributes. In *ICICS*, pages 141–152, 2007.
[17] Z. Liang and R. Wei. Efficient k-anonymization for privacy preservation. In *CSCWD*, pages 737 – 742, 2008.
[18] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, page 24, 2006.
[19] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, pages 223–228, 2004.
[20] M. R. Z. Mirakabad and A. Jantan. Diversity versus anonymity for privacy preservation. In *ITSim*, volume 3, pages 1–7, 2008.
[21] M. R. Z. Mirakabad, A. Jantan, and S. Bressan. Towards a privacy diagnosis centre: Measuring k-anonymity. In *CSA '08: Proceedings of the International Symposium on Computer Science and its Applications*, pages 102–107, Washington, DC, USA, 2008. IEEE Computer Society.
[22] M. E. Nergiz and C. Clifton. Thoughts on k-anonymization. *Data Knowl. Eng.*, 63(3):622–645, 2007.
[23] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer. From t-closeness to pram and noise addition via information theory. In *Privacy in Statistical Databases*, pages 100–112, 2008.
[24] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.
[25] T. M. Truta and A. Campan. K-anonymization incremental maintenance and optimization techniques. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, pages 380–387, New York, NY, USA, 2007. ACM.
[26] T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In *ICDEW '06: Proceedings of the 22nd International Conference on Data Engineering Workshops*, page 94, Washington, DC, USA, 2006. IEEE Computer Society.
[27] K. Wang, B. C. M. Fung, and P. S. Yu. Handicapping attacker's confidence: an alternative to -anonymization. *Knowl. Inf. Syst.*, 11(3):345–368, 2007.
[28] M. Wu and X. Ye. Towards the diversity of sensitive attributes in k-anonymity. In *IAT Workshops*, pages 98–104, 2006.
[29] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, pages 139–150, 2006.
[30] X. Xiao and Y. Tao. Personalized privacy preservation. In *SIGMOD Conference*, pages 229–240, 2006.
[31] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *SIGMOD Conference*, pages 689–700, 2007.
[32] X. Xiao and Y. Tao. Dynamic anonymization: accurate statistical analysis with privacy preservation. In *SIGMOD Conference*, pages 107–120, 2008.
[33] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu. Utility-based anonymization using local recoding. In *KDD*, pages 785–790, 2006.
[34] C. Yao, X. S. Wang, and S. Jajodia. Checking for k-anonymity violation by views. In *VLDB*, pages 910–921, 2005.
[35] Y. Ye, Q. Deng, C. Wang, D. Lv, Y. Liu, and J. Feng. Bsgi: An effective algorithm towards stronger l-diversity. In *DEXA*, pages 19–32, 2008.

[36] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing k-anonymization of customer data. In *PODS '05: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 139–147, New York, NY, USA, 2005. ACM.

Received 28 February 2010