Enhanced Additive Noise Approach For Privacy-Preserving Tabular Data Publishing

Saad A. Abdelhameed
Software Engineering and Information Technology Department
Faculty of Engineering and Technology,
Egyptian Chinese University
Cairo, Egypt
E-mail: shameed@ecu.edu.eg

Sherin M. Moussa
Information Systems Department
Faculty of Computer and Information Sciences, Ain Shams University
Cairo, Egypt
E-mail: sherinmoussa@cis.asu.edu.eg

Mohamed E. Khalifa
Vice President for Postgraduate and Research Studies
Egyptian Chinese University
Cairo, Egypt
E-mail: khalifa@ecu.edu.eg

***Abstract* - With the recent remarkable and fast evolution in telecommunication and computing technologies, great amounts of individuals' tabular-formatted data are collected and used by several organizations in the society. In some cases, some organizations need to share these gathered data to be used in business analysis, decision making or scientific researches purposes, which can involve sensitive information about the individuals. However, these data cannot be published in their original form to other third parties due to the associated privacy concerns. Consequently, preserving individuals' privacy represents a critical issue when sharing the individuals' private data. Hence, Privacy-Preserving Tabular Data Publishing (PPTDP) has received a great attention to protect the privacy of individuals' tabular data, where several approaches have been presented to address this issue. In this paper, we propose an enhanced additive noise approach for privacy-preserving microdata with Single Sensitive Attribute (SSA) publishing. The proposed approach maintains better published data utility to allow more accurate mining and analytical results, where more robust privacy protection against privacy attacks is provided.**

***Keywords* – data privacy; privacy-preserving data publishing; data anonymization; tabular data; single sensitive attribute; privacy attacks.**

I. INTRODUCTION

Many data holders have recognized the importance of their collected individuals' data for making business decisions, knowledge discovery, or other research needs [1]. These data holders include hospitals, government agencies, insurance companies… etc. However, these data often contain private and sensitive information about the individuals that must be well protected from being discovered through experiencing privacy disclosure attacks on the published data [2]. Hence, these data should be published in such a way that prohibits the disclosing of the individuals' identity by any intruders who want to breach the individual's privacy [3]. Thereby, data holders should publish data to the public without violating the confidentiality of personal information. This has raised major concerns about protecting the privacy of individuals while publishing data [2-5]. Therefore, Privacy-Preserving Tabular Data Publishing (PPTDP) has become an important research field in the recent years within the research community.

PPTDP studies how to transform the tabular data from its original version into a privacy-preserved form in order to be published to other parties. The main consideration of PPTDP approaches is to publish data with a strong privacy protection, such that the individuals' sensitive information could not be inferred with high confidence, while providing high utilization capabilities for useful mining and analysis tasks [2-3]. Such kind of transformation process is called the Anonymization process [4]. In general, PPTDP consists of three phases; the first is the data preparation phase, in which the data are collected and prepared by the data holder. The second is the data processing phase, in which these prepared data are processed and anonymized using a certain anonymization model. The third is the data publishing phase, in which the anonymized data table is published to the data recipients to be utilized in the desired analysis or research purposes.

Data anonymity model categorizes the data attributes into three types: (1) Explicit Identifier attributes (EI), which are the attributes that can typically identify an individual, i.e. name or social security number. (2) Quasi Identifier attributes (QIDs) that are not considered private individual data and can be recognized as background knowledge by other people or can exist in other external available databases, i.e. age and zip code. QIDs can potentially identify the individual if taken from the published data and linked together with such available data. (3) Sensitive Attributes (SAs) that are the private and unknown sensitive individual attributes, such as the disease and salary, which need to be prevented from being inferred and preserved against the different privacy disclosure attacks. Data anonymization rules do not publish EI, whereas QIDs may be masked using a certain disclosure control method, like generalization and/or suppression [6-7] or released without any treatment, which is known as bucketization method [8]. In the generalization, the QIDs values of the table's records are replaced by more general values according to a specific Domain Generalization Hierarchy (DGH) using either global or local recoding algorithms [6, 9]. In the published table, tuples having the identical generalized QIDs values are gathered together in a group called QI group or

978-1-5386-0821-0/17/$31.00 ©2017 IEEE

Equivalence Class (EC). In the bucketization, QIDs values are grouped properly into buckets according to the privacy principle imposed on the SA values. Thereby, the generalization approaches protect the tuples in the same EC from being disclosed using their QIDs values, while the bucketization approaches give more attention to offer better data utilization and less information loss.

In this paper, we propose the Enhanced Additive Noise (EAN) approach for PPTDP to anonymize and publish the static microdata with Single Sensitive Attribute (SSA). EAN enforces a newly-proposed privacy constraint on the SA value of the input tuple named "l-sensitive category diversity", whereas the QIDs original values are published to provide better data utility and attributes' distribution. In EAN, the input SA value of each tuple is replaced with a set of sensitive values consisting of the real SA value and $l-1$ random selected noise values, such that each sensitive value - either the actual one or the added noise ones - belongs to a different sensitive category. This cuts off any semantic relationship between the different sensitive values in the output sensitive set of the SA in each tuple. Therefore, the attacker will be prevented from disclosing the actual victim's SA value with confidence ratio higher than $1/l$. Besides, the attacker will not be able to infer any relationship or discover any additional information about the actual sensitive value of any input tuple using the added noise values. The rest of the paper is organized as follows. Section II reviews the related work. Section III defines the problem statement and the main contribution. In Section IV, we present the proposed EAN approach. Section V discusses the experimental results and the effectiveness of the presented method. Finally, section VI concludes the paper.

II. Related Work

Preserving the privacy of publishing static data with SSA has been considered by different research works in the recent years. The notion of data generalization is firstly introduced in [10] to achieve data anonymity when disclosing information for data privacy. Then, a privacy protection model named k-anonymity was proposed by L. Sweeney in [11]. It generalized the QIDs values to more general values, and then divided the records having the same generalized QIDs values into QI-groups, such that each QI-group contains at least k tuples. The main target of k-anonymity model is to make the anonymized tuples cannot be distinguished by their QIDs values using the generalization method. The k-anonymity model protected the published table against both the identity disclosure, which occurs when an individual is correctly identified by a certain tuple in the published table [2, 12-13], and the membership disclosure attack by which the presence or absence of an individual's tuple in the published table could be deduced [2-3, 13]. However, it failed to protect against the attribute disclosure, which occurs when new sensitive information of individuals are uncovered from the anonymized table [2-3, 14], the similarity attack, in which the SA values in an EC are similar or semantically related (i.e. the SA values belong to the same sensitive category) [2, 14-15], the skewness attack where the SA values in an EC are skewed to a specific value, and the sensitivity attacks in which the SA values are members of the same sensitivity level. This is due to the main drawback of the k-anonymity model, which is not enforcing any restrictions on the SA values in the anonymization process. Moreover, depending on generalizing each QID values has made it not suitable for high dimensional data, suffers from attributes correlations loss, and makes the published data lose significant information because of considering the distribution of each QID generalized values to be uniform, which is not true in the original table [2-3]. Accordingly, these limitations decrease the utility of the resultant anonymized data.

"p-Sensitive k-Anonymity" privacy model was proposed in [16] that obeys a restriction on the SA values in each QI-group to avoid the attribute disclosure attack issue of the k-anonymity. For each QI-group, it restricted the number of distinct values for each SA to occur at least p times within the same group. However, it could not avoid the attribute disclosure in some cases; i.e. p =1 (1-Sensitive 2-Anonymity), and p=2 (2-Sensitive 2-Anonymity). Besides, the presented restriction in "p-Sensitive k-Anonymity" failed to protect the published SA values against the similarity, the skewness and the sensitivity attacks. In [17], "l-diversity" generalization-based privacy model was presented depending on the principle of increasing the SA values' diversity in every QI-group, such that l number of sensitive values will be associated with each published tuple. The "l-diversity" model increased the difficulty to link the individual's tuple with a sensitive value with a probability ratio not higher than $1/l$, which made it more difficult with higher l values [18]. However, the model could not prevent the attribute disclosure in 2-diversity case. In addition, the SA values of the anonymized data with "l-diversity" model still face the skewness, the similarity and the sensitivity attacks.

Two enhanced privacy models: (p, α) sensitive k-anonymity and p+ sensitive k-anonymity were proposed in [19], extending the p-Sensitive k-Anonymity approach. The main focus of the new models was not on the SA values, but on the sensitive category that these values belong to and the weight by which these SA values contribute in each QI-group. The p+ sensitive k-anonymity model restricted the number of distinct categories for each SA to be at least p within the same QI-group. The (p, α) sensitive k-anonymity model enforced each QI-group to have at least p distinct SA values with their weight is at least α at total. Although the proposed models overcame the k-anonymity privacy attacks issues, but as being generalization-based approaches, they suffer from the curse of dimensionality and the remarkable information loss. In addition, the sensitive categories' similarity attack is issued in the p+ sensitive k-anonymity

model. Another privacy model called t-closeness was presented in [20], requiring that the distance between the distribution of a SA in any EC and the distribution of that SA in the overall table is no more than a threshold t, using The Earth Mover Distance (EMD) metric [21]. The defined distance among the SAs made the anonymized table with t-closeness model overcame the attacks of attribute disclosure, similarity and skewness. However, it is not suitable with various data tables with numerical SAs and it decreases the released data utility if such property is satisfied [22]. Besides, t-closeness property requires the SA values' distribution to be the same in any EC, which damages the correlations between the QIDs and SAs. In [23], (w,γ,k)-anonymity model was proposed based on the k-anonymity model, weight and the similarity of the SA values in order to protect the anonymized data against the different disclosure attacks. The presented (w, γ, k)-anonymity principle is satisfied in any EC if this EC satisfies the k-anonymity principle, its average weight is at least w and its similarity is at most γ. It successfully prevented the identity disclosure, attribute disclosure, similarity and sensitivity attacks on the anonymized data with both numeric and categorical SAs. However, it suffers from the curse of dimensionality, attributes correlation loss, information loss, long execution time due to relying on generalizing QIDs values.

Anatomy privacy model was introduced in [24], which published the exact QIDs values grouped into l-diverse buckets in a Quasi-Identifier Table (QIT) and the SA values with their counts in a Sensitive Table (ST). These separated tables were combined with a grouping mechanism based on the bucket (group) ID to construct the data tuple from the l-diverse published buckets. The main focus of Anatomy was to use bucketization method to maintain better data utilization capabilities than the generalization method-based approaches. This occurred as the exact QIDs-distribution of the original data table that is captured and reflected in the published QIT as well. Besides, Anatomy avoided the information loss resulting from the QIDs generalization. Therefore, it caused a significant less information loss in the anonymized data compared to the generalization-based approaches [25]. The division of the QIT and ST provided the privacy preservation guarantee through the difficulty for an adversary to deduce the actual SA value of a tuple from the SA l-diverse values in the same bucket. However, the release of the exact QIDs values makes Anatomy facing the identity and membership disclosure attacks. Moreover, the SA l-diverse values in the same bucket may not avoid the skewness, similarity and sensitivity attacks. Additionally, separating QIT and ST breaks the attribute correlations between the values of QIDs and SAs. Authors in [26] proposed the Permutation Anonymization (PA) model as an improved extension of Anatomy, which published the attributes values after being randomly permutated to increase the preserved privacy protection. This was by reducing the probability that an intruder can match all the QIDs values of a victim and then deduce the corresponding SA value from the l-diverse sensitive values in the same bucket compared to Anatomy. Consequently, PA still experienced the issued limitations of Anatomy. In [27], another data privacy model was proposed that worked through dividing the microdata into groups based on de-clustering the table tuples into groups according to their SA values. The de-clustering operation aimed to maximize the number of distinct SA values as possible in each EC. ECs were then formed with the QIDs values without generalization. This allowed each EC to have various records with distinct SA values as possible based on a dissimilarity function, which provided strong privacy preservation.

Ambiguity and PriView privacy models were proposed in [28] to protect the anonymized published data against both the membership and attribute disclosure attacks. Ambiguity published a corresponding table for each attribute of the QIDs containing its exact values unchanged, and a Sensitive Table (ST) containing the SA l-diverse values and their occurrence counts. PriView splits the original microdata into two tables only, each containing multiple QIDs to provide enhanced data utility and correlation between attributes. Although both models provided less information loss than generalization-based approaches, but Ambiguity caused more information loss than the other bucketization-based models, because of breaking the correlations between all the attributes. In addition, the SA l-diverse values in both models suffered from the skewness, similarity, and sensitivity attacks. Ambiguity+ model as an enhanced version of Ambiguity was proposed in [29] to retain better data utility. A new count column was added to the published tables in Ambiguity+ to prevent the uniform distribution assumption for the published tables' distinct values, in which more accurate data analysis was preserved rather than Ambiguity. However, Ambiguity+ provided the same privacy preservation level compared to Ambiguity.

Authors in [30] presented an Additive Noise (AN) privacy model to protect the published anonymized tables with the l-diversity principle against the intruders who may know some of the tuples' SA values as strong background knowledge. AN generalized the QIDs values of each input tuple in each QI-group and replaced its SA value by a sensitive value set consisted of its actual value and at least $(l-1)$ random selected noise values. This prevented an adversary from successfully disclosing the victim's SA value with a ratio higher than $1/l$ from the anonymized tables. Although AN maintained the frequency distribution of SA values from the original table, but it did not consider the case of the skewed SA values frequency distribution. Besides, satisfying the l-diversity principle among the sensitive values set was not sufficient to avoid the skewness, similarity and sensitivity attacks on these sensitive values. Table I summarizes the comparison between the different SSA privacy models in terms of the attacks they protect against.

SSA Privacy Models	Privacy Disclosure Attacks					
	Identity Disclosure	***Membership Disclosure***	***Attribute Disclosure***	***Skewness Disclosure***	***Similarity Disclosure***	***Sensitivity Disclosure***
k-anonymity	✓	✓	×	×	×	×
p-Sensitive k-Anonymity	✓	✓	✓	×	×	×
l-diversity	✓	✓	✓	×	×	×
p+ sensitive k-anonymity	✓	✓	✓	✓	×	×
(p, α) sensitive k-anonymity	✓	✓	✓	✓	✓	×
t-closeness	✓	✓	✓	✓	✓	×
(w, γ, k)-anonymity	✓	✓	✓	✓	✓	✓
Anatomy	×	×	✓	×	×	×
Permutation Anonymization (PA)	✓	✓	✓	×	×	×
De-clustering	×	×	✓	×	✓	×
Ambiguity, PriView and Ambiguity+	✓	✓	✓	×	×	×
Additive Noise Approach (AN)	✓	✓	✓	×	×	×

Table I. A comparison between the different SSA privacy models according to the attacks they protect against.

III. Problem Statement and Contribution

AN approach used the l-diversity principle to add noise values to the SA value of each tuple. This principle ensured the value diversity only among these sensitive values to be just different. However, it did not consider the relationships that may exist between these different sensitive values (i.e. some or all these values can be members of the same sensitive category or can be skewed to a certain value). This allowed the SA values of the anonymized tuple to face the skewness and similarity attacks, by which an adversary could acquire some information about the SA value with a high confidence ratio. For example, suppose a data table has an attribute "Salary" as a numerical SA, where an anonymized tuple can have the values (1K, 1.1K, 1.2K) as the sensitive values set satisfying the l-diversity principle (i.e. l=3). However, an attacker could deduce that the salary of this tuple's owner is relatively low or falls in the dense range [1K-1.2K] by the skewness attack.

Another instance, suppose a data table has an attribute "Diseases" as a categorical SA. Let an anonymized tuple has the values (Lymphoma, Heart attack, Leukemia) and another tuple has (Gastritis, Stomach cancer, Gastric ulcer) as the SA sensitive values set of each one, where both sets satisfy the l-diversity principle (i.e. l=3). However, an attacker could deduce that the first tuple's owner has a Cancer disease with a high probability of 2/3, whereas the second tuple's owner has a stomach-related disease with a high probability of 3/3 by the similarity attack. Besides, AN generalized the QIDs values of the anonymized tuples, which caused a valuable information loss, QIDs correlation loss and decreased the utility of the published data. These concerns represent critical limitations that should be addressed.

In this paper, our contributions can be summarized as follows. (1) We propose the Enhanced Additive Noise (EAN) approach for PPTDP to anonymize the static microdata with SSA. EAN relies on the noise addition concept that was employed in different previously-proposed approaches for privacy-preservation in data publication [3, 30-31]. (2) EAN enforces our newly-proposed privacy constraint on the SA values of the released tuples named "l-sensitive category diversity" that takes into consideration the semantic relationship between these sensitive values. Compared to AN, EAN efficiently protects against the attribute disclosure and the skewness and similarity attacks on the published SA sensitive values set, which improves the provided privacy level of the published tuples. (3) EAN releases the QIDs original values instead of generalizing them to avoid the associated information loss, maintain the correlation between QIDs, and to provide better data utility from the published table. (4) The sequential tuples processing method is used in EAN to anonymize all the input tuples individually instead of dividing them into groups. This is to protect against the attackers having strong background knowledge, like the real SA values of some tuples included in the original table. This also prevents the attackers from inferring a victim's real SA value if it is grouped with other tuples having the known sensitive values. Besides, this considers the case of the data table as well having a skewed frequency distribution of SA values. (5) Finally, the resultant information loss in the published tuples is considered as the data quality metric of the anonymization process in our proposed EAN, which was never discussed in AN approach. EAN utilizes the information loss metrics proposed in [32] that consider the information loss caused by only the inflation of SA value in each anonymized tuple in order to measure the utility of the published table.

IV. The Enhanced Additive Noise Approach

In this section, we introduce our proposed Enhanced Additive Noise (EAN) anonymization framework to preserve privacy of static data with SSA. First, the proposed data model is defined, followed by a detailed discussion for the proposed approach.

A. The Data Model

Let $T = \{QIDs, SA\}$ be the static data table that is needed to be published, where $QIDs = \{QID_1, QID_2, .., QID_n\}$ are the quasi identifiers, and SA is the sensitive attribute. For each

input tuple $t \in T$, we refer to $t.[QID_i]$ $(1 \le i \le n)$ as the value of QID_i of t, and to $t.[SA]$ as the sensitive value of SA. Let T^* be the anonymized published table of T, t^* is the anonymized tuple of t, and $t^* \in T^*$.

B. The proposed framework

EAN considers the data holder as the domain expert to categorize the distinct domain values of the SA into l different sensitive categories C_1, C_2, ..., C_x $(1 \le x \le l)$. Accordingly, EAN uses these values categorization with the guarantee of the "l-sensitive category diversity" restriction to add the noise values to the SA value of each input tuple. Then, the SA sensitive values set of each tuple is shuffled and published with the QIDs values in the anonymized tuple directly.

Definition 1. (l-sensitive category diversity) For each tuple $t^* \in T^*$, $t^* = (t.[QID_1], t.[QID_2], .., t.[QID_n], SV^*)$, SV^* is a randomly-sorted sensitive values set, representing the SA value of t^* in T^*. t^* is said to fulfill the l-sensitive category diversity principle if and only if each sensitive value in SV^* belongs to a different sensitive category C_x $(1 \le x \le l)$. T^* is said to fulfill the l-sensitive category diversity principle if each t_i^* $(1 \le i \le |T^*|)$ fulfills the l-sensitive category diversity. SV_1C_x is the first SA value in the sensitive category C_x, where $1 \le x \le l$. Fig. 1 represents an abstract architecture of the proposed EAN approach.

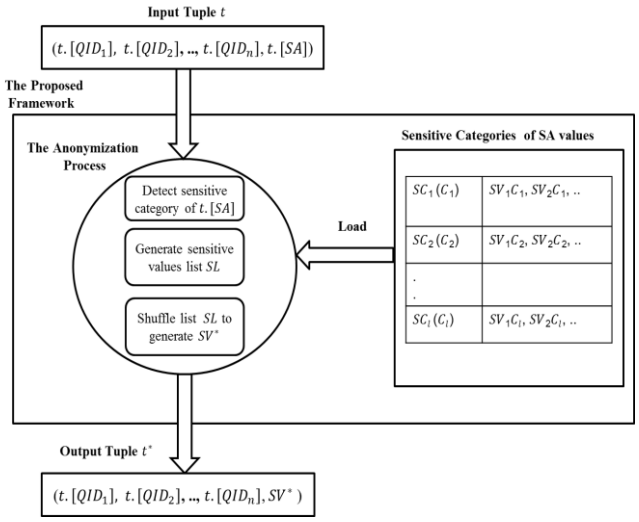

Fig. 1. An abstract architecture of the proposed EAN approach

1) The anonymization process

The anonymization process in EAN approach determines the sensitive category of $t.[SA]$ of each input t using the pre-defined sensitive values categorization. It then inflates $t.[SA]$ into a set of sensitive values SL, which obeys our proposed privacy restriction "l-sensitive category diversity". SL is shuffled to generate SV^*,and then SV^* is published with the exact QIDs values $t.[QID_1], t.[QID_2], .., t.[QID_n]$ as t^*.

2) Privacy preservation guarantees

The "l-sensitive category diversity" restriction adds exact $(l-1)$ counterfeit sensitive values to the real SA value in each tuple, such that each value will be associated with a different sensitive category. Thus, there will be no semantic relationship among all the sensitive set values of SA in t^*. Consequently, an adversary will not be able to deduce the actual victim's SA value with a confidence ratio higher than $1/l$. In addition, he will not be able to infer any relationship or discover any additional information about the actual sensitive value of any t^* using the added noise values. Thereby, EAN can efficiently protect against the attribute disclosure, skewness and similarity attacks in T^*.

3) Data utility measurement

EAN publishes the QIDs values of each t with their exact values instead of generalizing them in order to offer better data utilization from the published data. This would avoid the information loss of the generalization, which permits more effective data accuracy and analysis capabilities rather than AN, due to the capturing of the exact QIDs-distribution of T in T^* as well. In this proposed EAN approach, the information loss metrics presented in [32] are used, which consider the information loss caused by only the inflation of SA value in each anonymized tuple to measure the utility of the published table. These metrics are defined as follows:

$$IL(SA) = \frac{|l-1|}{l} \quad (1)$$

$$IL(t^*) = \frac{1}{n+1}\left[\sum_{i=1}^{n} IL(QID_i) + IL(SA)\right] \quad (2)$$

Where $IL(SA)$ is the SA information loss, l is the diversity parameter, n is the number of QIDs, $IL(QID)$ is the information loss of each QID, and $IL(t^*)$ is the total information loss occurred in each tuple t^*. Since EAN does not generalize QIDs values, $IL(QID_i)$ $(1 \le i \le n)$ is always equal to zero in each t^*.

4) The algorithm

EAN begins to work by loading the categorization file that categorizes the distinct domain values of the SA according to the domain expert. The value of diversity parameter l is dynamically determined according to this categorization as shown in Algorithm 1.

Furthermore, EAN reads the data table tuples sequentially. For each tuple t, EAN gets $t.[SA]$ and generates the randomly-selected sensitive values set including $t.[SA]$, such that $t.[SA]$ and each added value belong to a different sensitive category. Algorithm 2 shows the algorithm for the random set generation function. Accordingly, as shown in algorithm 3, this generated set is shuffled and assigned as the new SA value of t. The anonymized tuple t^* is then published, having the exact QIDs values of t and the new SA value that satisfy the "l-sensitive category diversity" restriction.

Algorithm 1. Load_sensitivecategories ($categorizationfile$)

Input: File $categorizationfile$
Output: sensitive-categories list senslist

1. *Create new list senslist; Let l = 0 and listindex = -1;*
2. *Lines =Readalllines (* $categorizationfile$*);*
3. ***For** each line in Lines*
4. *l =l+1; listindex = listindex+1;*
5. ***For** index = 0 to senslist.count*
6. ***If** index and listindex are equal*
7. *Insert (index, line) into senslist;*
8. ***End If**;*
9. ***End For**;*
10. ***End For**;*
11. *Return senslist;*

Algorithm 2. GenerateRandomSet ($t.[SA]$, *senslist*)

Input: SA value of t *and sensitive-categories list senslist*
Output: Sensitive random list SL_1

1. *Create new list* SL_1*;*
2. ***For** index = 0 to senslist.count*
3. ***If** senslist [index] does not contain* $t.[SA]$
4. *Select random value Svalue from senslist [index];*
5. *Insert Svalue into* SL_1*;*
6. ***End If**;*
7. ***End For**;*
8. *Insert* $t.[SA]$ *into* SL_1*;*
9. *Return* SL_1*;*

Algorithm 3. EAN (T)

Input: a data table T
Output: anonymized data table T^*
Precondition: EI and tuples with missing values are removed from T

1. ***For** each tuple* t *in T*
2. *Read* t*;*
3. *Create new list SL = GenerateRandomList* ($t.[SA], senslist$)*;*
4. *Shuffle SL;*
5. ***For** index = 0 to SL.count*
6. SV^**+= SL [index];*
7. ***End For**;*
8. *Assign* SV^* *to* $t.[SA]$*;*
9. *Publish* t *as* t^**;*
10. ***End For**;*

V. The Experimental Approach and Results

A framework has been developed using C# to evaluate our proposed EAN approach compared to AN's reported results in [30]. All experiments are performed on a machine with the same specifications as the discussed experiments of AN, which are 2.8 GHz Intel core processor with RAM of 2GB. The adult dataset from the UC Irvine Machine Learning Repository [33] was used in the experiments. The dataset consists of 32,561 tuples. The tuples with missing values were removed, where 30,162 valid tuples were used in our experiments. The same 8 attributes utilized in [30] were also selected in our experiments, which are the age, work class, education, marital-status, occupation, race, sex and country. The occupation is the SA and the other attributes are the QIDs. Several experiments were held to study the effect of changing the dataset size on the execution time of AN and EAN approaches, while fixing the diversity parameter l. Other experiments were dedicated to investigate how the execution time of both approaches is affected by varying the diversity parameter l, while fixing the dataset size. Besides, the resultant information loss per each tuple was also studied, while changing the diversity parameter l in EAN approach. All the parameters' values were set to be the same values as in AN's reported experiments [30] in order to settle the same testing environment of both approaches for a fair comparative evaluation. EAN used the sequential tuples processing method to anonymize all the input tuples individually. This makes the input tuples have a different set of sensitive values even if they have a similar SA value. Consequently, EAN takes into consideration the data tables having SA values with a skewed frequency distribution. Thus, our experiments can be categorized into three main categories as explained herein.

A. Execution time with different dataset sizes

The effect of changing the dataset size is studied with respect to the execution time of AN and EAN approaches, while fixing the diversity parameter. The diversity parameter is set to $l = 5$. As shown in Fig. 2 representing the execution time of both AN and EAN approaches, when the size of the dataset increases, the number of tuples needed to be anonymized increases. Thus, the execution time of both approaches increases.

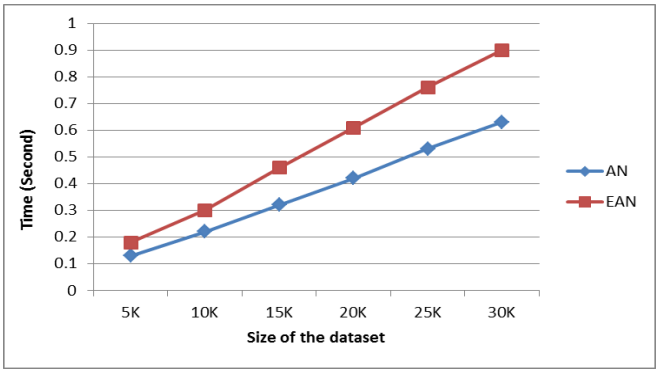

Fig. 2. Execution time for the different dataset sizes.

When the dataset size is 5K, AN consumes 0.13s to anonymize the tuples, while EAN consumes 0.18s with +0.05s representing 38.4% percentage increase. With the

dataset size of 10K, AN consumes 0.22s and EAN consumes 0.3s with +0.08s representing 36.3% percentage increase. In case of the 15K dataset size, AN consumes 0.32s and EAN consumes 0.46s with +0.14s representing 43.7% percentage increase. With the 20K dataset size, AN consumes 0.42s and EAN consumes 0.61s with +0.19s representing 45.2% percentage increase. When the dataset size was 25K, AN spends 0.53s to finish the anonymization, while EAN spends 0.76s with +0.23s representing 43.4% percentage increase. Finally, with the dataset size of 30K, AN spends 0.63s and EAN spends 0.9s with +0.27s representing 42.8% percentage increase. Thus, EAN needs an average of 41.6% additional execution time to anonymize the same dataset size, compared to AN. This is due to the additional restriction for choosing a random sensitive value that must belong to a different sensitive category from all other values in the same sensitive set for each t^*.

B. *Execution time with different diversity parameters*

The effect of changing the diversity parameter l is studied with respect to the execution time of AN and EAN approaches, while fixing the dataset size. The dataset size was set to 20K. As shown in Fig. 3, when the diversity parameter l increases, the number of the noise values needed to be added in each tuple increases. Thus, the execution time of both AN and EAN approaches increases. When the diversity parameter l= 3, AN consumes 0.32s to anonymize the tuples, while EAN consumes 0.42s with +0.1s representing 31.2% percentage increase. With l = 5, AN consumes 0.42s and EAN consumes 0.61s with +0.19s representing 45.2% percentage increase. In case of l = 7, AN consumes 0.61s and EAN consumes 0.8s with +0.19s representing 31.1% percentage increase. With l = 10, AN consumes 0.95s and EAN consumes 1.1s with +0.15s representing 15.7% percentage increase. Hence, the changing in the diversity parameter makes EAN consumes an average of 30.8% additional execution time compared to AN approach. This is due to the same additional restriction reason explained earlier for choosing a random sensitive value that must belong to a different sensitive category from all other values in the same sensitive set for each t^*.

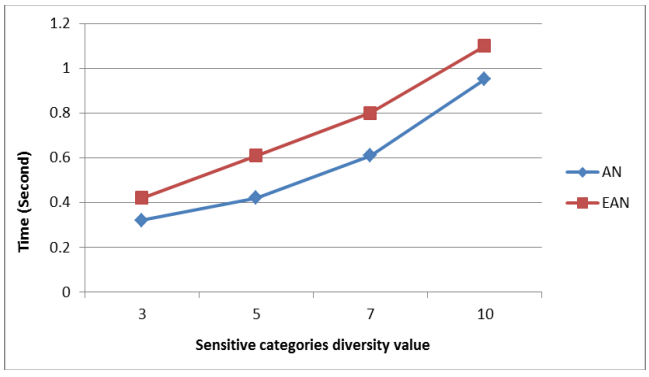

Fig. 3. Execution time with different diversity parameters

C. *Information loss per tuple with different diversity parameters*

In this experiment, the information loss occurred in each tuple due to the anonymization process is studied using the information loss metrics in (1) and (2). The experimentation of EAN approach is studied only, as the information loss per tuple was not considered in the results of AN reported in [30]. AN used the correlation loss metric to measure the utility of the anonymized data, which measures the SA and QIDs correlation in each divided group. Hence, it is not appropriate to be used with EAN, which does not partition the data table into groups. The results of EAN approach is shown in Fig. 4.

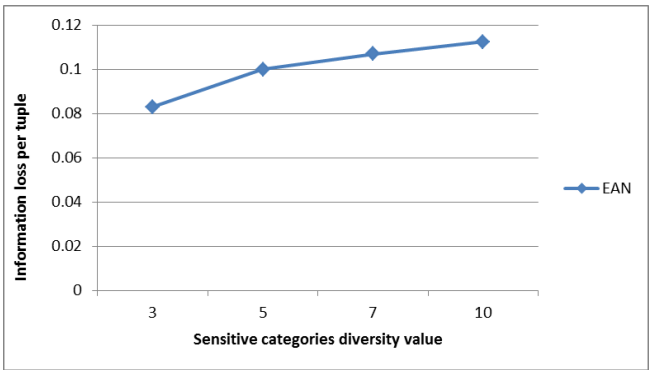

Fig. 4. Information loss per tuple with different diversity parameters.

As the diversity parameter l increases, more noise values are required to be added in each tuple. Thus, the information loss per tuple increases. When the diversity parameter l = 3, EAN incurred an information loss of 0.083 per each tuple. With l = 5, the resultant information loss per each tuple was 0.1. When l = 7, each anonymized tuple by EAN had an information loss of 0.107. When l= 10, EAN incurred an information loss of 0.1125 per each tuple. However, compared to AN, EAN incurs less information loss per tuple, as it does not generalize the QIDs values, which makes EAN avoid the generalization information loss on the contrary of that in AN.

VI. Conclusion

In this paper, the Enhanced Additive Noise (EAN) is introduced to anonymize the static microdata with Single Sensitive Attribute (SSA) with our newly-proposed "l-sensitive category diversity" privacy restriction. Each value in the sensitive values set in each anonymized tuple must belong to a different sensitive category. The experimental results show that EAN, compared to AN, consumes an average of 41.6% additional execution time to anonymize the same dataset size while fixing the diversity parameter, and an average of 30.8% additional execution time when the diversity parameter is changeable. This is due to the additional restriction for choosing a random sensitive value that must belong to a different sensitive category from all other values in the same sensitive set for each t^*. However,

EAN preserves better data utility and results in less information loss even with high l values. Besides, EAN efficiently protects against the attribute disclosure, skewness and similarity attacks on the SA values in the published data. Moreover, it considers the case of data tables having SA values with a skewed frequency distribution.

REFERENCES

[1] A.P. Singh and M.D. Parihar, "A review of privacy preserving data publishing technique", International Journal of Emerging Research in Management &Technology ISSN, pp. 2278-9359, 2013.

[2] S. A. Abdelhameed, S. M. Moussa, and M. E. Khalifa, "Privacy-Preserving Tabular Data Publishing: A Comprehensive Evaluation from Web to Cloud", Computers & Security, Elsevier, 2017, DOI: 10.1016/j.cose.2017.09.002.

[3] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments", ACM CSUR, 42(4), p.14, 2010.

[4] B. C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, "Privacy-preserving data publishing", Foundations and Trends® in Databases, 2(1–2), pp.1-167, 2009.

[5] V. S. Iyengar, "Transforming data to satisfy privacy constraints", In Proc. ACM SIGKDD, pp. 279-288. ACM, 2002.

[6] K. Wang, B. C. Fung, and S. Y. Philip, "Handicapping attacker's confidence: an alternative to k-anonymization", Knowledge and Information Systems, 11(3), pp.345-368, 2007.

[7] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity" In Proc. ACM SIGMOD-SIGACT-SIGART, pp. 223-228. ACM, 2004.

[8] W. Ya-Zhe, Y. Xiao-Chun, W. Bin and Y. Ge, "Privacy Preserving Approaches for Multiple Sensitive Attributes in Data Publishing [J]", Chinese journal of computers, 4, p.005, 2008.

[9] R.C.W Wong, J. Li, A.W.C. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing", In Proc. ACM SIGKDD, pp. 754-759. ACM, 2006.

[10] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information", In PODS, Vol. 98, p. 188, 1998.

[11] L. Sweeney, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), pp.557-570, 2002.

[12] P. Kiran, and N. P. Kavya, "A Survey on Methods, Attacks and Metric for Privacy Preserving Data Publishing", International Journal of Computer Applications, 53(18), 2012.

[13] Y. Xu, T. Ma, M. Tang, and W. Tian, "A survey of privacy preserving data publishing using generalization and suppression", Applied Mathematics & Information Sciences, 8(3), p.1103, 2014.

[14] N. Hamza, and H. A. Hefny, "Attacks on anonymization-based privacy-preserving: a survey for data mining and data publishing", Journal of Information Security, 4(02), p.101, 2013.

[15] N. Maheshwarkar, K. Pathak and V. Chourey, "Privacy Issues for k-Anonymity Model", International Journal of Engineering Research, Vol. 1, No. 4, 2011, pp. 1857-1861,2011.

[16] T.M. Truta, and B. Vinay, "Privacy Protection: p-Sensitive k-Anonymity Property", In ICDE workshops, p. 94, 2006.

[17] A. Machanavajjhala, J. Gehrke, and D. Kifer, "l-diversity: Privacy beyond k-anonymity". In Proc. ICDE'06, pp. 24-24. IEEE, 2006.

[18] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity", ACM TKDD, 1(1), p.3, 2007.

[19] X. Sun, L. Sun, H. Wang, "Extended k-anonymity models against sensitive attribute disclosure", Computer Communications 34, pp.526–535, 2011.

[20] N. Li, T. Li, S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity", In Proc. ICDE, pp. 106-115. IEEE, 2007.

[21] Y. Rubner., C. Tomasi, and L.J. Guibas, "The earth mover's distance as a metric for image retrieval", International journal of computer vision, 40(2), pp.99-121, 2000.

[22] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data", In Proc. ACM SIGMOD, pp. 473-486. ACM, 2008.

[23] X. Huang, J. Liu, Z. Han, and J. Yang, "A new anonymity model for privacy-preserving data publishing", China Communications, 11(9), pp.47-59, 2014.

[24] X. Xiao, Y. Tao, "Anatomy: simple and effective privacy preservation", In Proc. VLDB, pp.139–150, 2006.

[25] Y. Tao, H. Chen, X. Xiao, S. Zhou, and D. Zhang, "Angel: Enhancing the utility of generalization for privacy preserving publication", IEEE TKDE, 21(7), pp.1073-1087, 2009.

[26] X. He, Y. Xiao, Y. Li, Q. Wang, W. Wang, and B. Shi, "Permutation anonymization: Improving anatomy for privacy preservation in data publication.", In Proc. PACKDDM, pp. 111-123. Springer Berlin Heidelberg, 2011.

[27] Q. Wei, Y. Lu, and Q. Lou, "Privacy-Preserving Data Publishing Based on De-clustering." In ICIS, Seventh IEEE/ACIS, pp. 152-157. IEEE, 2008.

[28] H. Wang, "Privacy-preserving data sharing in cloud computing", Journal of Computer Science and Technology, 25(3), pp.401-414, 2010.

[29] M. Rajaei and M.S. Haghjoo, "An improved Ambiguity+ anonymization technique with enhanced data utility", In Proc. IKT, pp. 1-7. IEEE, 2015.

[30] H. Zhu, S. Tian, M. Xie, and M. Yang, "Preserving Privacy for Sensitive Values of Individuals in Data Publishing Based on a New Additive Noise Approach", ICCCN, pp. 1-6. IEEE, 2014.

[31] R. Brand, "Microdata protection through noise addition", in Inference control in statistical databases: Springer, pp. 97-116, 2002.

[32] S. Kim, M.K. Sung, and Y.D. Chung, "A framework to preserve the privacy of electronic health data streams", Journal of biomedical informatics 50, pp.95-106, 2014.

[33] http://archive.ics.uci.edu/ml/index.php