Accepted Manuscript

Title: Privacy-preserving tabular data publishing: a comprehensive evaluation from web to cloud

Author: Saad A. Abdelhameed, Sherin M. Moussa, Mohamed E. Khalifa

PII: S0167-4048(17)30184-0
DOI: http://dx.doi.org/doi: 10.1016/j.cose.2017.09.002
Reference: COSE 1198

To appear in: *Computers & Security*

Received date: 18-3-2017
Revised date: 2-9-2017
Accepted date: 6-9-2017

Please cite this article as: Saad A. Abdelhameed, Sherin M. Moussa, Mohamed E. Khalifa, Privacy-preserving tabular data publishing: a comprehensive evaluation from web to cloud, *Computers & Security* (2017), http://dx.doi.org/doi: 10.1016/j.cose.2017.09.002.

Privacy-Preserving Tabular Data Publishing: A Comprehensive Evaluation from Web to Cloud

Saad A. Abdelhameed[1], Sherin M. Moussa[2], and Mohamed E. Khalifa[3]
[1,3] Faculty of Engineering and Technology, Egyptian Chinese University (ECU), Cairo 11351 Egypt
[2] Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566 Egypt
[1]shameed@ecu.edu.eg, [2]sherinmoussa@cis.asu.edu.eg, [3]khalifa@ecu.edu.eg

Authors' Biography

Saad A. Abdelhameed received the Bachelor's degree in Information Systems from the Faculty of Computer and Information Sciences (FCIS), Ain Shams University (ASU), Egypt in 2012. He is now a Master student at FCIS, ASU and teaching assistant at the Egyptian Chinese University (ECU), Egypt. His research interests include Information Security, Privacy Preservation and Big Data Privacy and Publishing.

Sherin M. Moussa is associate professor at FCIS, ASU. She had her PhD April 2010 in conjunction with the University of Illinois, Urbana-Champaign, USA. In addition to her teaching, research contributions and international publications, she has +17 years of experience in IT multiple/complex national/regional projects. Her research interests are Big Data Analytics, Data Mining, Mobile Data Management, Data Streaming, GIS, Software Engineering, and Big Data Privacy and Publishing. Dr. Moussa was the Director of the Quality and Accreditation Unit, Foreign Affairs Consultant at the Research & Development of Projects Center, and Director of Grants and International Collaboration Office at ASU.

Mohamed E. Khalifa is professor of Mathematics. He is president of Egyptian Chinese University. He was Dean of FCIS-ASU, Dean of Future Academy, and IT Consultant of Benha University. Prof. Khalifa got his PhD in Mathematical Modeling & Simulation, U.S.S.R, 1980. He has +34 years of experience in local/international educational management, research, and teaching. He was in scientific mission to Courant Institute of Mathematical and Computer Sciences, and visiting Professor in New York University. He had major contributions in justice to resolve IT-related legal cases. His research interests include Bioinformatics, Image Processing, Data Security, and Big Data Privacy and Publishing.

***Abstract* -The amount of data collected by various organizations about individuals are continuously increasing. This includes diverse data sources often for data of high dimensionality. Most of these data are stored in tabular format and can include sensitive content. Preserving data privacy is an essential task in order to allow such data to be published for different research and analysis purposes. In this context, Privacy-Preserving Tabular Data Publishing (PPTDP) has drawn considerable attention, where different approaches have been proposed to preserve the privacy of individuals' tabular data. Such data can include Single Sensitive Attributes (SSA) or Multiple Sensitive Attributes (MSA) or come from data streams. In this paper, we conduct a comprehensive study to analyze and evaluate the main different data anonymization approaches that have been introduced in PPTDP. The study investigates the three broad areas of research: SSA, MSA and data streams. A detailed criticism is presented to highlight the strengths and the weaknesses of each approach including their deployment in the cloud and Internet of Things (IoT) environments. A research gap analysis is discussed with a focus on capturing current state of the art in this field in order to highlight the future directions that can be considered.**

***Keywords* – data privacy; privacy-preserving data publishing; data anonymization; data streams; multiple sensitive attributes; single sensitive attribute.**

I. INTRODUCTION

NOWADAYS many enterprises that are actively collecting and storing individuals' data from numerous sources into large databases have recognized the potential value of these data as an important information source for making business decisions and researches [1]. The persistently increasing amounts of such data make their privacy a challenging and vital task, especially when the data are highly dimensional. In general, privacy preservation concerns are related to authentication, data accessing, data encryption, and data publishing. Many data holders need to publish their microdata for different purposes in such a way that does not disclose the individuals' identity. Thus, Privacy-Preserving Data Publishing (PPDP) has been given a considerable measure of attention in the recent years within the research society.

PPDP studies have been conducted on tabular and graph data. This paper focuses on researches on Privacy-Preserving Tabular Data Publishing (PPTDP) in relation to privacy preservation when publishing tabular data only. PPTDP investigates the transformation of the original tabular data when being published to other parties into a privacy-preserved version that protects the privacy of the records' owners and their sensitive information from being disclosed while still providing high utility. This transformation process is known as the anonymization process [2]. Data anonymity represents a protection model that allows sharing of sensitive and private data with a guarantee that the individuals - the subjects of the data - cannot be recognized, while the privacy-preserved published data should remain valuable and utilized to support effective data analysis and research tasks. Generally, PPTDP has three phases: data preparation, data processing and data publishing phases. The data publisher collects and prepares the data to be processed and anonymized. Finally, the processed / anonymized data table is sent to the data recipients for further analysis or research purposes. Fig. 1 shows an abstract architecture of PPTDP.

The differential privacy and partition-based models are the main directions of the privacy preservation concepts that are commonly used in the PPTDP field [3]. Differential privacy [4-5] is a privacy model that makes no assumption about an adversary's background knowledge [3, 6]. It uses a randomized mechanism that ensures the probability of any released output of queries' responses is equally likely from all nearly identical datasets, making an individual's privacy independent of whether his record is included in the dataset or not [7-8]. The differential privacy publishing scenario can be either interactive [9] or non-interactive [10]. In the interactive settings, the data owner must answer the received query from the data miner before the next query is issued and responded to. In the non-interactive settings, the data owner receives all the queries at one time, and then releases their responses. The responses to the queries may be modified by adding noise to preserve privacy, whereby the strong privacy guarantees come at the price of noise added to each query response. On the other hand, the partition-based privacy models divide a data set into groups using different anonymization techniques. The adversary's background knowledge is carefully taken into consideration, whereas the privacy is ensured by imposing some constraints on the released data (i.e. the output record is required to be indistinguishable among k records or the sensitive value is well-represented in each group). Many approaches have been introduced in this regard. Hence, the partition-based privacy models are the focus of this study.

The anonymity model in the partition-based privacy approaches divides the data into three types of attributes: (1) Explicit Identifier attributes (EI) which represent the very specific and personal data that can distinctively identify an individual, such as the name or social security number. (2) Quasi Identifier attributes (QIDs) which are concerned with the non-private / sensitive data of an individual that may be known to other people as background knowledge or may be available in other publicly databases that can potentially identify the individual variables if taken and linked together, such as age, birth date, gender and zip code. (3) Sensitive Attributes (SAs) which are the private data of an individual that are unknown to the others, such as the disease and salary. In particular, SAs are the attributes that an intruder wants to infer and to discover from the published data. Accordingly, these three types of attributes need to be well-protected against the privacy-related disclosure attacks. Data anonymization regulations forbid the release of EI whereas QIDs are masked using different disclosure control methods in order to ensure that no adversary can correctly infer the SAs of persons.

Three main areas are considered; (i) anonymizing static data with a Single Sensitive Attribute (SSA), (ii) anonymizing static data with Multiple Sensitive Attributes (MSA), and (iii) privacy preservation in data streams. The most popular anonymization methods are the generalization-based and bucketization-based methods. The generalization-based methods work through different operations, either generalization, suppression or both. Generalization replaces the QIDs values of the original data table's records by more general values according to a given taxonomy using either the global or the local recoding algorithms [11]. Suppression replaces and suppresses some of the QIDs values by a special value, i.e. '*', through either value suppression [12] or local suppression (cell suppression) [13]. These records, having the same generalized QIDs values, are then grouped together in a group called QI-group or Equivalence Class (EC) in the published table. In addition to generalizing the QIDs values, some generalization-based approaches can also obey a certain restriction on the SAs values. These approaches are referred to as generalization-based with restricted sensitive values approaches. The bucketization-based approaches publish the exact values of the QIDs without generalization, and then separate between the QIDs and SAs in the published table [14]. Hence, the generalization-based approaches aim to protect the tuples in the same EC from being distinguished by their QIDs values,

while the bucketization-based approaches are concerned with maintaining better data accuracy and utilization. Table I presents an example for an original raw data format in PPTDP.

As shown in Table I, the records represent the information about some patients in a medical dataset that needs to be published and where the attribute {Name} is an EI, the attributes {Age, Gender, Zip code} are the QIDs, and the attribute {Disease} is the SA. On the other hand, the tremendous scaling of the individuals' data streams in many critical applications creates a crucial necessity for data anonymization [15]. Current anonymization approaches applied on the static data are NP-hard, which makes them inapplicable for the real-time processing of data streams to preserve privacy [16]. Thus, various techniques have been investigated to anonymize data streams, taking into consideration their speed of generation. Fig. 2 shows the deduced categorization for the different PPTDP approaches.

In this paper, a detailed analysis and evaluation study are provided for the main data anonymization approaches that have been investigated in the PPTDP field, starting from the web publishing stage to the usage stage of the cloud and Internet of Things (IoT) environments. Our study considers the static data and data streams for both SSA and MSA. The strengths and weaknesses of each approach are highlighted, where a research gap is conducted to evaluate the current state position in this research field and to indicate the future directions that could be considered. The rest of the paper is organized as follows: section 2 presents the various PPTDP approaches applied to the static data with only SSA; section 3 discusses the PPTDP approaches considered for the static data with MSA. Section 4 studies the PPTDP approaches for the data streams while section 5 investigates the PPTDP approaches in the cloud environment. Section 6 investigates the PPTDP approaches in the IoT environment; section 7 discusses the current research gap in the PPTDP field and deduces the possible future directions in this field. Finally, section 8 concludes the paper.

II. Single Sensitive Attribute Approaches

Several researches have been conducted to preserve the privacy of static data with SSA. Perturbation-based approaches were investigated as a statistical disclosure control, which replace the original individuals' data values or the results of queries by some artificial data values by either swapping [17], condensation [18], adding noise [19], or micro-aggregation [20]. Perturbative methods ensure that the statistical analysis resulted from these perturbed data does not vary significantly from that computed from the original data. However, the main drawback of such category of privacy preservation methods is that the data integrity is impaired and damaged. Hence, the perturbed published data records are considered "synthetic", as they do not correspond to the original data representing the real individuals. Therefore, the individual records in the perturbed data have no meaning to the recipients [15]. In contrast, the non-perturbative methods, including the generalization-based and bucketization-based approaches, generate less precise data that are semantically consistent with the raw data. Thus, the truthfulness of the published data is preserved. Such approaches can be categorized as follows:

A. Generalization-based Approaches

P. Samarati and L. Sweeney first introduced the concept of data generalization in [21] to provide data anonymity when disclosing information in order to preserve data privacy. L. Sweeney then proposed the first formal privacy protection model named k-anonymity [22]. This model depends on generalizing all the values of QIDs with more generic values, and then dividing the records having the same QIDs values into groups called QI-groups. The altered microdata table T fulfills k-anonymity property if every combination of the QIDs in T occurs at least k times, making the likelihood ratio to correctly distinguish an individual to be at most $1/k$.

The k-anonymity model secures against the identity disclosure, also known as the "record linkage attack" [15, 23-24]. This happens when an individual is effectively perceived by a specific record in the published table, so that the attacker can uniquely identify the victim's record in the published table. The k-anonymity model secures as well against the membership disclosure attack, also known as the "table linkage attack" [15, 24], which allows any intruder to know whether the published dataset includes an individual's record or not. This gives the attacker a confident knowledge about the presence or absence of a victim's record in the released table. However, this model is not suitable for high dimensional data, where each generalized value is always an exceedingly wide interval. In addition, it loses the correlations between the different data table attributes because each attribute is generalized separately, which represents an obstacle to the efficient analysis of the attribute correlations. Moreover, the generalization of QIDs to more general values leads to valuable information loss in the published data, because of the data uniform-distribution assumption expected by the researcher when answering a query compared to the original data. Information loss is considered as one of the main data quality metrics of the anonymization process, which is used to measure the usefulness and utilization of the published data. The detection of valuable information loss negatively impacts the published table utility for research and analysis purposes.

Besides, this model does not impose any restrictions on the SA values in the published data. Thus, it could not protect against the attribute disclosure, also known as the "attribute linkage attack" [15, 24], which happens when new sensitive information about some individuals is inferred and uncovered from the published data that increases the confidence of an adversary to infer the SA of a certain victim from the published data table. This can occur by linking the QIDs of the published table data with any external available data tables or the adversary's background knowledge. Some researches refer to the attribute disclosure or the identity disclosure as the "linking attack" [15]. Accordingly, the k-anonymity model is seriously vulnerable to the similarity attack, also known as the "homogeneity attack" [15, 24-26], where the sensitive values in a QI-group are similar and lack for diversity or semantical similarity (i.e. the different sensitive values belong to the same sensitive category). In addition, there is the skewness attack, in which the sensitive values in a QI-group are skewed to a certain value, and the sensitivity attack, where the different sensitive values belong to the same sensitive level in the published tables. Table II represents an example for 3-anonymimty table of Table I.

As shown in Table II, the attribute {Name} is removed and the values of the QIDS {Age, Zip code} are generalized, producing three QI-groups with the following generalization schemas <[20, 30], M, [16k, 25k]>, < [30, 45], F, [30k, 55k]>, <[50, 60], F, [60k, 75k]>, including tuples 1-3, 4-6 and 7-9 respectively. For instance, the age 27 and zip code 16k of tuple 1 have been replaced in the first QI-group by the intervals [20, 30] and [16k, 25k] respectively. The age 42 and zip code 54k of tuple 5 have been replaced in the second QI-group by the intervals [30, 45] and [30k, 55k] respectively. The age 57 and zip code 62k of tuple 8 have been replaced in the third QI-group by the intervals [50, 60] and [60k, 75k] respectively.

B. Generalization-based with Restricted Sensitive Values Approaches

Several approaches have been induced as a variation to the k-anonymity method. T.Marius and B.Vinay presented a new privacy model "p-Sensitive k-Anonymity" [27]. It aims to avoid the attribute disclosure problem of the k-anonymity model by applying a certain restriction to the sensitive values in each QI-group. The altered microdata table T is considered fulfilling the p-sensitive k-anonymity property if it satisfies the k-anonymity property, and for each QI-group, the number of distinct values for each SA occurs at least p times within the same group. However, this model still faces the attribute disclosure problem in some cases; i.e. p=1, 2-sensitivity 2-anonymity, where in case of p=1 (1-sensitivity 2-anonymity), at least 50% of the tuples will have the same SA value within the same EC, while in p=2 (2-sensitivity 2-anonymity), 100% of the tuples will have the same SA value within the same EC. This allows the intruder to infer the SA of the victim's tuple with high confidence ratio. Besides, this model may not avoid the similarity attack, because it is possible that an EC may contain most or all the p distinct SA values belonging to the same pre-defined sensitive category, in addition to the sensitivity attack when most or all these sensitive values may belong to the same pre-defined sensitivity level. For example, the disease sensitive values (esophagus cancer, leukemia, lymphoma, lung cancer and stomach cancer) belong to the Cancer sensitive category, as they are semantically-related. The data holder can define these same sensitive values into different sensitivity levels, i.e. (esophagus cancer, lung cancer and leukemia) belong to the sensitivity level 1 and (stomach cancer and lymphoma) belong to the sensitivity level 2. In [28], an improved privacy model was presented depending on a different principle called "l-diversity" to overcome some shortcomings of the k-anonymity model. This principle increases the diversity of the SA values in every QI-group, so that each tuple will be associated with l sensitive values. A modified microdata table T is considered satisfying the l-diversity property if each QI-group contains at least l "well-presented" values for the SA. Two metrics were used to measure the utility of these anonymized data, which were the Discernibility Metric (DM) [29] and the Normalized average EC size metric (CAVG) [30]. These metrics are defined as follows:

$$Discernibility\ Metric\ (DM) = C_{DM} = \sum_{E} |E|^2 \tag{1}$$

$$Normalized\ average\ EC\ size = C_{AVG} = \frac{|T|}{numE \cdot k} \tag{2}$$

Where $|E|$ is the EC size, k is the anonymity variable, $|T|$ is the total number of records, and $numE$ is the number of all ECs. This model enhances the difficulty of linking a sensitive value to an individual into a confidence ratio not higher than $1/l$, making it so difficult with higher l values. However, the model may be insufficient to prevent the attribute disclosure in some cases; i.e. 2-diversity case, in which 50% of the tuples have the same sensitive value in each QI-group, allowing an adversary to infer the sensitive value of a record with high confidence of 0.5. Besides, it suffers from the skewness, similarity and sensitivity attacks. The skewness attack occurs when the l-diverse SA values are distinct in the same EC, but these values are skewed or have very dense proximity to a certain value, specifically in case of the numerical SAs. As for the similarity attack, it occurs

when the l-diverse SA values are distinct in the same EC, but most or all of them may belong to the same pre-defined sensitive category (semantically-related) or there may be one value appear much more frequently than other values within the same EC, making it easy for an adversary to conclude that a record in that EC is very likely to have that value. As for the sensitivity attack, it occurs when the l-diverse SA values are distinct in the same EC, but most or all of them may belong to the same pre-defined sensitivity level. Thus, the l-diversity privacy requirement is not considered as a sufficient restriction to prevent these attacks. Another weakness of l-diversity is that achieving high l values may be difficult in many microdata. Being a generalization-based approach, the published data table suffers from losing considerable information compared to the original one, which decreases the data analysis accuracy of the published data. Table III represents an example for the 2-diversity table of Table I.

As shown in Table III, the generalization produced two QI-groups with the generalizations <[20, 30], M, [16k, 25k]> and <[30, 60], F, [30k, 75k]>, including tuples 1-3 and 4-9 respectively. Each QI-group contained at least 2 different sensitive values of the SA disease, i.e. the first QI-group contained hepatitis and flu as sensitive values, and the second QI-group contained gastritis, leukemia, stomach cancer, heart disease and HIV. For instance, the age 27 and zip code 16k of tuple 1 have been replaced in the first QI-group by the intervals [20, 30] and [16k, 25k] respectively. The age 42 and zip code 54k of tuple 5 have been replaced in the second QI-group by the intervals [30, 60] and [30k, 75k] respectively.

X. Sun, L. Sun and H. Wang proposed in [31] two new enhanced privacy models: (p, α) sensitive k-anonymity and p+ sensitive k-anonymity as an extension of the p-sensitive k-anonymity approach. They focused on the sensitive category that the values belong to rather than the specific values of SAs in order to overcome the drawbacks of k-anonymity, especially the attribute disclosure. The p+ sensitive k-anonymity approach sorts the values of the sensitive categorical attributes according to their sensitivity, forming an ordered value domain, and then partitions the attribute domain into x categories and obtains the QI-groups. A modified microdata table T is considered satisfying the p+ sensitive k-anonymity property if it satisfies the k-anonymity property, and for each QI-group in T, the number of distinct categories for each SA is at least p within the same QI-group. The (p, α) sensitive k-anonymity model still considers the specific values of SAs, but it includes a weight to measure how much the values of a SA contribute in each QI-group, formulated as follows:

Let $D(S) = \{S_1, S_2, \dots, S_m\}$ denotes a partition of the categorical domain of an attribute S and let weight (S_i) denotes the weight of category S_i. Then,

$$\begin{cases} Weight\ (S_i) = \dfrac{i-1}{m-1}\ ;\ 1 \le i < m \\ Weight\ (S_m) = 1, \end{cases} \tag{3}$$

The total weight of any QI-group is the summation of weights of each sensitive value included in the QI-group. Hence, an altered microdata table T is considered fulfilling the (p, α)-sensitive k-anonymity property if it satisfies the k-anonymity, and each QI-group has at least p distinct SA values with their total weight at least α. The data quality metrics presented in equations (1) and (2) were used to measure the utility of the anonymized data. However, the two proposed approaches have some shortcomings, like the curse of dimensionality and the valuable information loss, since they are generalization-based approaches. In addition, p+ sensitive k-anonymity model faces the similarity attack at the sensitive categories' level. This is because an EC with at least p number of distinct categories for the SA may contain many sensitive values belonging to the same certain category which enables an adversary to infer the sensitive category of the record's sensitive value in this EC with a high confidence ratio. A novel privacy model called t-closeness was proposed in [32]. This model requires that the distribution of an SA in any EC is close to the distribution of this attribute in the overall table. The Earth Mover Distance (EMD) metric was employed to measure the distance between the two distributions [33]. A QI-group is considered to fulfill the t-closeness necessity if the distance between the distribution of an SA in this group and the distribution of that attribute in the whole table is no more than a threshold t. The data table is supposed to have t-closeness if all the QI-groups ensure t-closeness.

The data quality metrics presented in equations (1) and (2) were used as well to measure the utility of the anonymized data. The t-closeness model overcame the problems of SAs disclosure, similarity and skewness attacks issued in the k-anonymity and l-diversity by defining a semantic distance among the SAs, which can adequately diminish the amount of the individual's data that an adversary can gain from the released table. However, this may not be appropriate with various data tables, especially those having numerical SAs, which may require a computational process to apply this property. Moreover, the data utility is greatly degraded if such a process is obtainable, because enforcing t-closeness damages the correlations between the QIDs and SAs due to the requirement of having the distribution of the sensitive values to be the same in each QI-group.

In [34], another privacy model named (n, t)-closeness was presented based on the t-closeness model. An EC E_1 has (n, t)-closeness if there exists a set E_2 of records that is a natural superset of E_1 such that E_2 includes at least n records, where the

distance between the two distributions of the SA in E_1 and E_2 does not exceed a threshold t. A table has (n, t)-closeness if all the ECs have (n, t)-closeness. The "natural superset" in (n, t)-closeness is as the reference class used in [35]. Another data utility metric was presented to measure the information loss of the anonymized data using the information loss of an EC and the entropy of the SA values in the EC defined as follows:

$$IL(E_1, \dots, E_p) = \sum_{1 \le i \le p} \frac{|E_i|}{|T|} H(E_i) \tag{4}$$

Where T is the original dataset, E_i $(1 \le i \le p)$ is an EC in the anonymized data, $H(T)$ is the entropy of the SA values in T, $H(E_i)$ is the entropy of the SA values in E_i $(1 \le i \le p)$, $IL(E_1, \dots, E_p)$ is the total information loss of EC and $U(E_1, \dots, E_p)$ is the utility of the anonymized data. While the (n, t)-closeness model better secured and improved the utility of the released data, but the EC with t-closeness or (n, t)-closeness may cause a proximity breach in a table of numeric SAs [36]. Therefore, both models still have the risk of the sensitivity attack. H.Xuezhen, J.Liu, Z.Han and J. Yang presented in [37] a new model named (w,γ,k)-anonymity to protect against the identity disclosure, similarity and sensitivity attacks in the anonymous data based on the k-anonymity model. Any EC satisfies the (w,γ,k)-anonymity if it satisfies the k-anonymity, its average weight is at least w and its similarity is at most γ. The table satisfies the (w,γ,k)-anonymity if every EC in the table satisfies the (w,γ,k)-anonymity, where the EC similarity is:

$$\gamma_E = 1 - d_E \tag{6}$$

d_E is the minimum record-EC distance within the EC, named as the EC separation, which is defined as:

$$d_E = min_{u \in E}\, d(u, E) \tag{7}$$

Where $d(u, E)$ is the distance between a record and the corresponding EC and named record-EC distance, which is based on the mean value of distances between this record and all the other ones in that EC such that:

$$d(u, E) = \sum_{\substack{v \in E \\ v \ne u}} \frac{d(u,v)}{|E|-1} \tag{8}$$

Where $|E|$ represents the size of the EC and $d(u, v)$ is the distance between the two SAs values u and v, which is defined to be:

$$d(u, v) = \frac{level(u,v)}{H} \tag{9}$$

Where $level(u, v)$ is the height of the lowest common ancestor of u and v, and H is the height of the SAs taxonomy tree. The data quality metrics presented in equations (1) and (2) were also used to measure the utility of the anonymized data. The proposed model effectively prevented the similarity, sensitivity, attribute disclosure and identity disclosure attacks on the anonymized data. In addition, this model can be applied for both numeric and categorical SAs, as its three parameters have definitions on both types.

C. Bucketization-based Approaches

X. Xiao and Y. Tao presented "Anatomy" in [38] as a novel model, in which all the quasi-identifiers and sensitive values were released in two separate tables: Quasi-Identifier Table (QIT) and Sensitive Table (ST) combined with a grouping mechanism to overcome the data utilization weaknesses of the generalization-based approaches. Anatomy first creates the QIT based on l-diversity groups so that each tuple in each QI-group in the QIT includes all its exact QIDs values, together with its group membership in a new column Group-ID. It then produces the ST, which retains the statistics of the SA values of each QI-group in a new column count.

The Anatomy model allowed more effective data utilization, accuracy and analysis than the generalization-based publication methods, due to the capturing of the exact QIDs-distribution of the original data table in the published QIT. It guaranteed privacy preservation through the division of the QIT, ST and the used grouping method. However, Anatomy is vulnerable to the identity and membership disclosure attacks. This is due to the release of the QIDs exact values. Besides, the SAs values in the published ST may face skewness, similarity and sensitivity attacks, where the ST obeys the l-diversity privacy requirement that is not a sufficient restriction to prevent these attacks as explained earlier. Other shortcomings are revealed when the number of the recurring sensitive value in the microdata is so huge; the number of distinct sensitive values in each QI-group will be decreased, making it easy to infer the correct sensitive value. In addition, it breaks the attribute correlations between the QIDs and SAs by separating them in the two tables QIT and ST. Tables IV(A) and IV(B) represent the anonymized tables of Table I using the

Anatomy model, where Table IV(A) is the Quasi-Identifier Table (QIT) and Table IV(B) is the Sensitive Table (ST) with 2-diversity.

As shown in Tables IV(A) and IV(B), the tuples are divided into three QI-groups, including tuples 1-3, 4-6 and 7-9 respectively. Each QI-group contained at least 2 different sensitive values of the SA disease, i.e. the first QI-group contained hepatitis and flu as sensitive values, the second QI-group contained gastritis, leukemia and stomach cancer and the third QI-group contained leukemia, heart disease and HIV. The QIT included the exact QIDs values of the records, with Group-ID column indicating to which QI-group the record belongs to. The ST included the SA values of each QI-group with their statistics in the count column. For instance, the first two tuples of the ST indicate that the two tuples of the first QI-group are associated with Flu, and one tuple with Hepatitis.

The Permutation Anonymization (PA) model was presented in [39] as an improved version of Anatomy, compared to the generalization-based approaches that resulted in huge information loss [38, 40]. This model produces two tables as in Anatomy. While Anatomy directly publishes all the QIDs-values without additional treatment, PA publishes the attributes values after random permutation. This provided stronger privacy preservation guarantees, allowing an intruder to have less likelihood to deduce the sensitive value of a victim compared to Anatomy on a similar microdata partition. The Normalized Certainty Penalty (NCP) was used to measure the information loss [41]. NCP showed that PA retains significantly more information in the microdata rather than Anatomy and provides good data utility, allowing highly effective data analysis. Thus, PA can be applied to many real applications, such as the Location-Based Services (LBS), in which all the attributes are sensitive, in addition to the applications where membership attack is a critical concern [42, 43]. However, the SAs values in the published data may face skewness, similarity and sensitivity attacks even after the permutation, where the ST obeys the l-diversity privacy requirement that is not a sufficient restriction to prevent these attacks as explained earlier. Another model was proposed in [44] to preserve data privacy by partitioning the microdata into groups based on de-clustering. The technique de-clustered the records into groups according to their sensitive values, such that the number of distinct sensitive values was as large as possible in each group. QI-groups were then obtained containing the exact QIDs values of the records without generalization. In the de-clustering operation, the record is assigned to the group of the highest dissimilarity based on a certain distance function that measures the dissimilarity between two records $r_i$and r_j, defined as follows:

$$Distance(r_i, r_j) = |r_i[A^s] - r_j[A^s]| \tag{10}$$

Where A^s is the SA value, $r_i[A^s]$ is the SA value of record i and $r_j[A^s]$ is the SA value of record j. This allowed the QI-group to contain as much different number of records and distinct sensitive values as possible to guarantee strong privacy preservation. Average Protecting Expectation (APE) was introduced to measure the degree of privacy protection, which considered the number of records in addition to the number of distinct sensitive values involved in each QI-group, rather than the number of distinct sensitive values only as in l-diversity, formulated as follows:

$$Pr_{\text{APE}} = \prod_{i=1}^{n}(1 - \frac{1}{l_i} \cdot \frac{|QI_i|}{N}) \tag{11}$$

Where $|QI_i|$ is the number of records contained in QI_i, l_i is the number of distinct sensitive values in QI_i, n is the number of the QI-groups and N is the total number of records in the microdata. However, this method consumes extra time for the record assigning process to a certain group. It also suffers from the identity disclosure and membership disclosure attacks, as it does not generalize the QIDs values. Moreover, SAs may face the similarity and sensitivity attacks, when most or all the distinct SA values in the same group belong to the same pre-defined sensitive category (semantically-related), or to the same pre-defined sensitivity level respectively. H. Wang in [45] proposed "Ambiguity" and "PriView" models to protect against both the membership and attribute disclosures in the published anonymized microdata with low information loss. Ambiguity publishes the exact values of QIDs in separate tables. For each one of QIDs, it releases a corresponding table and a Sensitive Table (ST) with the sensitive values and their frequency counts. In order to provide better data utility and attributes correlation, PriView approach splits the original microdata into only two tables, each containing multiple QIDs. The relative error of count queries formula was used as the information loss metric for both techniques, which is defined as follows:

$$Error = \frac{|Q(T) - Q(T^*)|}{Q(T)} \tag{12}$$

Where Q is a count query, $Q(T)$ and $Q(T^*)$ are the accurate and approximate result by applying Q on the original table T and the released table T^* respectively [46]. Ambiguity provided less information loss than generalization, while PriView incurred

information loss less than Ambiguity. However, Ambiguity has worse information loss rather than the other bucketization-based approaches, as it breaks the correlations between all the attributes in the published data. In addition, the SAs values in both Ambiguity and PriView face similarity, sensitivity and skewness attacks, where the ST in both models follow the l-diversity privacy requirement that is not a sufficient restriction to prevent these attacks as explained earlier. Authors in [47] proposed Ambiguity+ model that published the frequency of each distinct value to preserve better data utility as an improvement of Ambiguity. Ambiguity+ model added a count column to the published tables in order to decrease obscurity in the published data and to prevent the uniform distribution assumption for the distinct values of the published tables. This increased the analysis accuracy and preserved better data utility rather than Ambiguity, while privacy was still carefully maintained. Ambiguity+ has no impact on privacy compared to Ambiguity. Therefore, SAs values still face similarity, sensitivity and skewness attacks. Table V summarizes the attacks facing the main SSA privacy models, while Table VI provides a detailed comparison between the different SSA privacy models.

III. MULTIPLE SENSITIVE ATTRIBUTE APPROCHES

Considering real applications that deal with high-dimensional SAs, many studies have been directed towards the privacy preservation of static data with MSA. The contributions can be categorized as follows.

A. Generalization-based with Restricted Sensitive Values Approaches

Slicing model was proposed in [48] which partitions the data table both horizontally and vertically. Vertical partitioning groups the attributes into columns, where each column contains a subset of the table's attributes that are highly correlated. Horizontal partitioning gathers the tuples into buckets (groups) that satisfy the l-diversity principle, and then randomly switches the values in each column in each bucket. Slicing prevented the membership disclosure by partitioning the attributes into more than two columns, as well as the attribute disclosure through the random permutation of the values in each l-diverse bucket. It preserved the data utility since it grouped the highly-correlated attributes together in addition to the correlations between the QIDs and SAs while handling the high-dimensional data. However, Slicing still has a drawback when different tuples have the same QIDs and sensitive values, which may give the same original tuple while performing the random permutation process. Besides, the SAs values are vulnerable to the skewness, similarity and sensitivity attacks, since they follow the l-diversity privacy requirement that is not a sufficient restriction to prevent such attacks as clarified earlier. Mondrian Slicing and Suppression Slicing were presented in [49] as two enhanced slicing models to overcome the drawback of Slicing. In Mondrian Slicing, the random permutation was done for all the buckets, not within a single bucket as in Slicing. As for Suppression Slicing, any of the QIDs values in a tuple was suppressed if it did not satisfy the l-diversity privacy requirement. It then performed the Slicing. Therefore, Suppression Slicing maintained the data utility with the minimum loss by suppressing only very few of the QIDs values, while preserving the privacy by the random permutation. However, both models enhanced the data utility capabilities rather than the privacy guarantees of the Slicing model.

B. Variations on Generalization-based with Restricted Sensitive Values Approaches

F.Luo, J.Han, J.Lu and H.Peng proposed in [50] an improved framework, named "ANGELMS" (Anatomy and Generalization on Multiple Sensitive attributes). It vertically partitioned the attributes into several SAs tables and one QIDs table. In each table, the tuples were then divided into groups (buckets). As for the SAs tables, the tuples in each sensitive bucket were allocated to obey the l-diversity requirement. In the QIDs table, the QIDs values of each QI-group were generalized under the k-anonymity principle. An additional information loss metric and suppression ratio were proposed to measure the quality of the anonymized data as follows:

$$AddInfoLoss = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{b_j}\frac{|G_i| - l}{b \cdot l} \tag{13}$$

$$SuppRatio = \frac{n_s}{n} \tag{14}$$

Where G_i is a group in an l-diverse table, b is the number of groups, m is the number of SAs, n_s is the number of suppressed tuples and n is the total number of tuples. ANGELMS prevented the attribute disclosure and overcame the identity and membership disclosure problems faced by the anatomy technique by generalizing the QIDs values, but it has some weaknesses. It breaks the attribute correlations between the QIDs and the SAs, since it vertically partitions the attributes into several SAs tables and a QIDs table preventing the efficient analysis of attributes' correlations. In addition, it may lose considerable information being a generalization-based approach. The Additive Noise model was proposed in [51] to publish the anonymized tables under the l-diversity principle against the attacker who has strong background knowledge about the published data. This approach replaces the SAs value of each record by a value set consisted of its actual SAs value and at least $(l-1)$ random

selected noise values, making the attacker unable to disclose the victim's SA value from the published table with a certainty higher than $1/l$. The Normalized correlation Loss Penalty (NLP) and GLP as a normalized version of NLP were used to measure the correlation loss of the anonymized table, defined as follows:

$$NLP(t^*) = \sum_{s \in (S - S_t)} P_r(t^*[SA] = s) \tag{15}$$

$$GLP(T^*) = (\sum_{t^* \in T^*} NLP(t^*)) \div |T| \tag{16}$$

Where T^* is the anonymized table of data table T, t^* is the anonymized record of t that belongs to the same individual, S is the SA values set of T, s is a SA value, $t^*[SA]$ is the SA value of t^*, and S_t is the SA values set consisting of the SA values of the records having the same Non-Sensitive Attributes (NSAs) values with t. $GLP(T^*)$ showed that the proposed model provided less correlation loss than that of the generalization-based, Anatomy and Slicing models by having smallest values, while preserving the frequency distribution of all SAs values obtained from the published table to be the same as that of the original table. However, it does not handle the skewness attack, in which the frequency distribution of SAs values in the data table is skewed, in addition to the similarity and sensitivity attacks. The similarity attack occurs when the l-diverse values of the sensitive list in each record are distinct, but most or all of them may be semantically-related, i.e. belonging to the same pre-defined sensitive category. On the other hand, the sensitivity attack occurs when the l-diverse values of the sensitive list in each record are distinct, but most or all of them may belong to the same pre-defined sensitivity level.

In [52], (p^+)-sensitive, t-closeness model was proposed, combining the advantages of the t-closeness and the p-sensitive k-anonymity approaches to overcome the similarity and skewness attacks of the anonymization techniques [27, 32]. As clarified earlier, the t-closeness requirement considers a QI-group fulfilling the t-closeness requirement if the distance between the distribution of an SA in this group and the distribution of that attribute in the whole table does not exceed a threshold t. The data table satisfies t-closeness if all the QI-groups satisfy t-closeness. Whilst, the p^+sensitivity is used to generate p distinct sensitivity levels of SAs values in each QI-group. Accordingly, (p^+)-sensitive, t-closeness model applies the t-closeness concept on the sensitivity levels, where the sensitivity level of SAs is determined and disseminated in the QI-groups in such a way that each QI-group has at least p distinct sensitivity levels of the attribute values under the defined threshold t. Thereby, it preserved the SAs values. P-cover k-anonymity model was proposed in [53] by extending the Incognito algorithm [54]. It protected against the identity, membership, positive and negative MSA disclosures through the fulfillment of the proposed MSA-P-Diversity principle among the associated MSA in the published table. This principle stated that the attacker should eliminate no less than $P - 1$ sensitive values to successfully disclose the sensitive value of a record in each QI-group in the published MSA table. A taxonomy tree-based metric was presented as an enhanced data quality metric [29, 55]. It captured the actual amount of data distortion as the height of the generalized data values, defined as follows:

$$Distortion\ (T) = \sum_{i=1}^{n} \sum_{j=1}^{m} |h_{ij}| \tag{17}$$

Where $T = \{t_1, \dots, t_n\}$ is the microdata table with $QIDs = \{Q_1, \dots, Q_m\}$, and $|h_{ij}|$ is the height of the generalized value if the value in Q_i of t_i has been generalized $|h_{ij}|$ levels up in the taxonomy tree. The main drawback is that satisfying the MSA-P-Diversity principle when the number of SAs is large will significantly increase the information loss in the released table, which is considered a serious limitation.

C. Bucketization-based Approaches

In [56], Y.Ye, Y.Liu, D.Lv, and J.Feng presented the Decomposition model, where it decomposed the data table into SA-groups and each group contained exactly l distinct sensitive values. The tuples were then grouped properly, with their QIDs values unchanged instead of generalizing them. Therefore, the tuples within the same QI-group will share the union of their sensitive values, in order to maximize the number of such SA-groups as much as possible to maintain stronger privacy.

The Decomposition model was also applied on MSA, where one SA donated as the "primary sensitive attribute" was chosen then the technique formed the SA-groups according to it, such that the original values of each SA-group and each non-primary sensitive attribute were combined. The duplicated values were counted once, reducing the privacy disclosure risk. The Decomposition model maintained the attributes correlation and allowed more effective data utilization, as the QIDs values remained unchanged instead of generalizing them. In addition, it protected against the similarity and attributes disclosure attacks faced by the k-anonymity-based publishing approaches, introducing a new MSA diversity principle ($(l_1, l_2, \dots, l_d)$ - diversity). This principle stated that each SA-group G_i and each $i \in \{1,2, \dots, d\}$ contains at least l distinct sensitive values, so that the tuples within a QI-group share the union of their sensitive values, maximizing the number of such SA-groups. However, this model is

vulnerable to the identity disclosure and membership disclosure attacks, as it does not generalize the QIDs values, in addition to the skewness and sensitivity attacks. The skewness attack is faced when the SAs values in each SA-group are distinct. However, these values may be skewed or have very dense proximity to a certain value, specifically in the numerical SA cases whilst the sensitivity attack is faced when the SAs values in each SA-group are distinct, but most or all of them belong to the same pre-defined sensitivity level. Table VII summarizes the attacks facing the main MSA privacy models, whereas Table VIII provides a detailed comparison between the MSA privacy models.

IV. Data Stream Approaches

Recent studies have started to target privacy preservation for the tabular data streams. Some researches depend on generalizing the QIDs using either the specialization tree or clustering, while others use the bucketization method, in which the exact QIDs values are released using the count-based or time-based sliding windows approaches. The count-based sliding windows approach requires waiting to accumulate a certain number of tuples in order to start the anonymization process, whereas the time-based sliding windows approach adjusts the size of the sliding windows in terms of time in order to ensure efficient streaming data anonymization via the adaptive resizing methodology. The following sub-sections categorize the main efforts presented in this field.

A. Generalization-based Approaches

J.Li, B.Chin and W.Wang presented in [57] a model named SKY (Stream K-anonYmity) that applied k-anonymity on data streams with a delay constraint using the specialization tree for privacy protection. The SKY algorithm starts the specialization tree from the root node and then grows with the data streams. It can also construct the specialization tree through applying an offline algorithm on the historical stream data. To periodically adjust the specialization tree, it divides the tree nodes into two classes; (i) work nodes, which are used to publish at least k tuples to satisfy the k-anonymity property, and (ii) candidate nodes that waits to satisfy the k-anonymity property accordingly. While reading a tuple t from the input stream, SKY inspects the specialization tree to locate the most specific generalization node g that covers t. If g is a work node, then tuple t is anonymized with g and is out directly. Otherwise, if g is a candidate node, t is stored into its frequency set $FS(g)$ until the k-anonymity property or the delay δ-constraint is satisfied to save the tuple from being expired. A general information loss metric was adapted for the data streams to measure the anonymization information loss. Considering a categorical attribute, given a value v in its Domain Generalization Hierarchy (DGM), the information loss of the value v is defined as follows:

$$VInfoloss\ (v) = \frac{|S_v| - 1}{|S| - 1} \tag{18}$$

Where S_v is the set of leaf nodes of the sub-tree rooted at v in the DGH and S is the set of leaf nodes in the same DGH. As for the continuous attributes, given a value interval $I = [l, u]$ from domain $[L, U]$, its information loss is defined as follows:

$$VInfoloss(I) = \frac{u - l}{U - L} \tag{19}$$

Accordingly, the information loss of any generalization $g = < v_i, \dots, v_m >$ will be equal to:

$$Infoloss\ (g) = \frac{1}{m}\sum_{i=1}^{m} VInfoloss(v_i) \tag{20}$$

Although the technique protects against the identity disclosure, it directly applies the k-anonymity model. This makes the published streaming data suffer from the previously-mentioned shortcomings of the k-anonymity model. In addition, it imposes more time overhead in the search process of the specialization tree for a suitable generalization node to publish the arrived tuples, which is unacceptable in the streaming environment. On the other hand, the numerical values anonymization using the specialization tree increases the difficulty of the anonymization process, due to the difficulty of finding a suitable hierarchy, as well as making the published data vulnerable to be re-distinguished if an adversary discovers the used generalization hierarchy.

CASTLE (Continuously Anonymizing STreaming data via adaptive ClustEring) model was presented in [58]. It anonymized the data streams by defining clusters as n-dimensional intervals in the QIDs domains that satisfied a delay constraint δ to ensure the freshness of the anonymized data. For every newly-arriving tuple, CASTLE checks whether a tuple t in some cluster will expire, satisfying the delay constraint. It works through two main cases; the first is when the size of cluster C, hosting the expiring tuple, is already greater than or equal to k, that is, C is a k-anonymized cluster. In this case, CASTLE outputs all the tuples in C with C's generalization and keeps it to be reused for anonymizing other expiring tuples. The second case is when cluster C has a size less than k, that is, C is a non-k-anonymized cluster. CASTLE verifies if t can fall in a k-anonymized

cluster. If true, t is directly released with its generalization. Otherwise, CASTLE performs merge operations between C and a non- k-anonymized cluster C_i to become a k-anonymized cluster, which brings the minimum enlargement to C among all the other non- k-anonymized clusters. It then outputs t with C's generalization and considers just the non- k-anonymized clusters as the possible candidates to accommodate the new-arrival tuple.

CASTLE protects against the identity disclosure and uses a cluster-based scheme with reusable clusters principle that provides less time consuming compared to using the specialization trees. However, it does not restrict the maximum number of tuples in a cluster, which leads to the linear growth of the cluster size according to the data stream size. This causes high information loss, as being a generalization-based method. In addition, CASTLE still suffers from the attribute disclosure, sensitivity, similarity attacks, and attributes correlations loss. This is because CASTLE adopts the k-anonymity privacy model, which also has no certain restriction on the SA values in the published data. Since CASTLE splits the resulted clusters that have a size more than $2k$ tuples, waiting for at least k tuples for each cluster to publish its tuples, it significantly increases the time complexity of the approach, which is inconvenient with data streams. In [59], B-CASTLE was proposed as an extended cluster-based scheme of CASTLE that restricts the maximum number of tuples to α in each cluster, computed as follows:

$$\alpha = \left\lfloor \frac{\delta}{\sqrt{k+\beta}} \right\rfloor \tag{21}$$

Where δ is the threshold of the maximum publishing delay deadline, k is the number of tuples needed by a cluster to publish its tuples and β is the maximum number of clusters generated by CASTLE. B-CASTLE merges few parts of the non-k-anonymized clusters that contain the tuples going to expire at a time for publishing instead of merging them all as in CASTLE, limiting the maximum release delay of each tuple to improve the approach total time complexity. However, B-CASTLE still suffers from the remaining weaknesses of CASTLE mentioned earlier.

Weak Clustering-based Data Streams k-Anonymity (WCDSA) framework was presented in [60] for data stream publishing. It generalized the stream tuples using an extended clustering feature tree concept of BRICH algorithm [61-62]. WCDSA anonymized the data streams under δ-delay constraint with minimal information loss of anonymous data streams. The algorithm clustered the data stream tuples into different clusters, then checked for clusters that satisfied the k-anonymity property. The minimal information loss cluster was then generalized and output directly. This dynamically updated the clustering feature tree with the arrival of data streams and production of its anonymous outputs. The size of each cluster was restricted between k and $2k-1$ in order to optimize WCDSA algorithm [16]. The generic information loss metrics presented in equations (18), (19) and (20) were used to measure the information loss of each generalized cluster in the clustering feature tree [63]. The proposed method prevented the attribute disclosure in the data streams publishing process. However, this method resulted in high information loss and suppressed tuples ratio especially in the cases of larger k with smaller δ delay time, which damaged the published data utility and usefulness. This is due to its checking for the cluster size just at the time when the total number of tuples is an integral multiple of $2k-1$, instead of checking it at the time each tuple comes.

B. Bucketization-based Approaches

Authors in [64] presented SANATOMY (Stream ANATOMY), a data streams privacy preserving publishing model based on Anatomy [38]. It applied the l-diversity principle on the stream tuples' sensitive values under a δ-constrained publishing strategy. SANATOMY works through two main processes; the first is the buckets generating process, in which the stream tuples are dynamically partitioned into l-diverse buckets. The second is the tuples publishing process, in which a bucket with a tuple to be published is anonymized using the Anatomy model into QIT and ST based on δ-constrained strategy. The tuple is then appended into the stream quasi identifiers table (S_{qit}) and stream sensitive table (S_{st}) respectively. If the bucket does not meet the l-diversity principle, SANATOMY merges it with the l-diverse bucket one by one till the l-diversity is satisfied. If it still cannot be published, it re-partitions all the buckets' tuples into new buckets, then anatomizes and outputs them.

Compared to CASTLE, the buckets generation process requires time upper-bounded by $O(b)$, where b is the number of buckets kept in memory. The publishing process requires time upper-bounded by $O(n)$, where n is the number of non-published tuples, whereas the space complexity is also upper-bounded by $O(n)$. Besides, SANATOMY performs well when the data streams have a speed of 15 tuples/second, whereas CASTLE performs poorly even when the speed is up to 9 tuples/second. Meanwhile, SANATOMY improves the published QIDs precision, applies the l-diversity principle on the sensitive data, reduces time complexity, and retains considerable information for efficient data analysis. However, it suffers from the membership disclosure and identity disclosure attacks, as it publishes QIT including all its exact QIDs values. This is in addition to the extra time overhead, due to the merging and re-partitioning processes of the buckets that do not meet the l-diversity principle.

A Delay-Free (DF) anonymization framework was proposed in [65] for preserving the privacy of electronic health data streams in real time based on the idea of Anatomy technique. DF inflates each SA of the input tuples into l-diverse values. When a tuple t arrives, DF generates a Quasi-Identifier Tuple (QIT) and an inflated Sensitive Tuple (ST) and publishes them directly. The QIT is produced from the QIDs in the tuple t, whereas the ST is an l-diverse set of artificial sensitive values including the original one. QIT and ST can be joined by the join key groupID. The DF anonymization algorithm is divided into two parts: late

validation and counterfeit generation. When a tuple arrives, DF tries to find a present counterfeit value accepted by the tuple per a provided condition. If it exists, the tuple is published as QIT with the groupID of the counterfeit, and the tuple is updated with an increased count value. Otherwise, DF generates l-diverse counterfeit sensitive values. In order to measure the quality of DF approach, an information loss data utility metric was presented that considers only the inflation of sensitive values, since the exact QIDs values are released without distortion. This metric was defined as follows:

$$IL_a(a_{sens}) = \frac{|ST_j - 1|}{|ST_j|} \tag{22}$$

Where $|ST_j|$ is the number of distinct values of all sensitive values, including the counterfeits and the original value in a group. Accordingly, the information loss of tuple t, considering the sensitive attribute, is measured by applying equation (22) in the generic information loss metric presented in equation (20) as follows:

$$IL(t) = \frac{1}{n+1}\left[\sum_{i=1}^{n} IL_a(a_i) + IL_a(a_{sens})\right] \tag{23}$$

The proposed approach does not generate a significant delay during the anonymization process, as it releases the tuple at a time without waiting to accumulate a certain number of tuples. It preserves data privacy with a probability of $1/l$ by fulfilling the l-diverse requirement on the sensitive values. However, SAs values in the published tables are vulnerable to the skewness, similarity and sensitivity attacks, as the l-diversity principle is not considered as a sufficient restriction to prevent these attacks as discussed earlier. In addition, the publishing of data with inflated sensitive values list containing unreal and counterfeit values can decrease the utility of the anonymized results. Moreover, releasing the exact QIDs values makes the published data suffer from the identity disclosure and membership disclosure attacks.

C. Count-based Sliding Window Approaches

SWAF (Sliding Window Anonymization Framework) model was proposed in [66] to ensure k-anonymity of data streams. It uses the Sliding window (Sw) concept, which is a buffer that maintains the most recent part of a data stream and replaces the oldest tuple by a new one during the continuous arrival of the data stream tuples. SWAF deals with the Sw as a static dataset and executes SK algorithm to construct the specialization tree. In addition, it uses IK algorithm to continuously adjust the specialization tree while the Sw is being updated, ensuring that no node violates the k-anonymity property. Updating Sw includes inserting the new tuple into the Sw and deleting the oldest one from it. Let g be the most specific generalization node and $FS(g)$ is its frequency set. If $FS(g) = |k-1|$, then the node does not satisfy the k-anonymity property. However, SWAF directly anonymizes the data streams using the Sw under the k-anonymity principle, which makes the published results suffer from the k-anonymity shortcomings. In addition, it uses the specialization tree, which causes two main problems; (i) It imposes more time overhead with the search process for a suitable used generalization node to publish the arrived tuples. (ii) It increases the difficulty of the numerical values anonymization, due to the difficulty in finding an appropriate generalization hierarchy, which makes the published data vulnerable to be re-identified if an adversary discovers the used hierarchy. Moreover, the merging process of nodes not satisfying the k-anonymity property leads to extra time overhead, which is not appropriate for data streams.

H.Zakerzadeh and S.Osborn presented in [67] a cluster-based k-anonymity model called FAANST, which anonymized numerical streaming data using the Sw processing. It uses three parameters: k, MU and $DELTA$, where k represents the minimum number of tuples needed by a cluster to publish its tuples, MU is the total number of tuples in the Sw representing the size of the processing window and $DELTA$ is the data loss threshold. In each round, FAANST waits for MU tuples to arrive, then partitions them into M clusters using the k-means algorithm, such that $M = size\ of\ current\ window/k$. It then checks if the number of tuples in each cluster c_i is $\geq k$, that is, the cluster fulfills the k-anonymity requirement. If the condition is satisfied, FAANST outputs all the tuples in c_i, while keeping the other tuples to the next rounds and storing the clusters that fulfill the cluster size $\geq k$ and info loss $\leq DELTA$ that will be reused in the next rounds to output tuples. Once the number of tuples in the Sw achieves MU again in the next rounds, the algorithm checks if the tuples fall into one of the accepted clusters found in previous rounds. If so, they are output to that cluster directly and re-apply the algorithm on the remaining tuples. FAANST has several drawbacks; it neither supports categorical data nor restricts a certain deadline for the delay that can each tuple tolerate. Thus, a tuple may remain in the system longer than the delay time constraint, causing time-sensitive records being processed to be expired. Besides, it waits for the window to accumulate a specific number (MU) of tuples to go to the next round, resulting in more time overhead.

The same authors presented in [68] two delay-sensitive passive and proactive models as extensions for FAANST. The two approaches applied a user-defined delay threshold for the time that every tuple can remain in the system, forcing each tuple to be output once it exceeds this certain time delay. In the passive solution, the arrival time for each tuple is saved in each round and the non-expired tuples are published with the released tuples by the original clusters of FAANST. The remaining non-output

tuples are suppressed if they have passed their deadline. The proactive solution goes one step more by estimating the time the tuple will spend in the next round until it will be revisited to be anonymized using a simple heuristic. This determines whether a tuple will expire if it is not released in this round. This heuristic supposes that the next round will take $(currentTime - timeOfLastVisit)$, where $timeOfLastVisit$ represents the time of the last visit for each tuple. Accordingly, the proactive approach publishes the tuple in the current round if it will expire. Otherwise, this tuple is kept for the following round to be output to a cluster having a better accommodation and information loss. The factors of FAANST, in terms of the average waiting time of each tuple and the number of expired tuples, were improved by the proposed methods. However, the verification of record expiration causes an additional overhead that negatively impacts the response time. Besides, a record can get repeatedly processed and recycled till it expires. In addition, the window size still waits to accumulate a specific number (MU) of tuples to go to the next round, resulting in extra time overhead while still supporting numerical attributes only.

A parallel anonymization model named FAST was proposed to anonymize big data streams using a multi-thread technique [69]. A proactive time-expiration heuristic was applied as follows to publish data before they expire:

$$\big((\,current\ time - Arrival\ time) + estimated\ round\ time\ < expiration\ time\big) \tag{24}$$

In the first round, FAST continuously reads δ tuples and passes them to new threads until the number of threads reaches to a specific threshold. Each thread then publishes its set of tuples, ensuring that the k-anonymity requirement is applied. FAST selects the closest $k-1$ tuples to the first tuple t by calculating the distance between the tuples, and then inserts them into a new cluster. This cluster is then generalized and saved in the list of reusable clusters to best cover and publish additional tuples, where the estimated round time is updated at the end of each round. Based on the proactive time-expiration heuristic, other tuples remaining in the set are kept for processing in another round or published immediately with the appropriate cluster. In the next rounds, tuple t is published with the reusable cluster (Ck_best) generalization if (Ck_best) covers t with the smallest information loss. Otherwise, tuple t and its neighbors are published with another new cluster C_{new} generalization. The information loss metrics presented in equations (19) and (20) were used to measure the information loss of each one of the QIDs to determine the information loss rate of each cluster, defined as follows:

$$infoloss(c, G) = \omega \times infoloss(G(c).QID) \tag{25}$$

Where c is the cluster, G is the generalization function, $\omega = \omega_i \dots \omega_n$ is a weighted vector of size $n \cdot 1$, ω_i is the weight of the i^{th} quasi-identifier attribute and $\sum_1^n \omega_i = 1$. Moreover, a cost function was defined to show the latency impact in the data publication as follows:

$$Cost(c, G) = infoloss(c, G) \times (1+\alpha)^{(PublishedTime - arrivalTime)} \tag{26}$$

Where α is the latency parameter, $arrivalTime$ and $PublishedTime$ are the times when the data arrived and published respectively. This model improved the big data stream anonymization in terms of information loss and latency, thanks to the applied parallelism concept. However, it still uses a generalization-based method for anonymizing tuples that causes valuable information loss, in addition to the count-based method that requires waiting for at least (δ) tuples to start the anonymization. This increases the time complexity, which is inconvenient with data streams. Besides, the published data suffer from the attribute disclosure, similarity and sensitivity attacks, because the k-anonymity privacy model is adopted, which has no certain restriction on the SA values in the published data.

D. Time-based Sliding Window Approaches

In [70], a novel privacy-preserving scheme was proposed on data streams using the k-anonymity model and buffer size adjustment based on the data arrival rate. The time-based Sw concept was applied instead of the count-based Sw to handle the delay problem in data streams k-anonymization schemes, which wait for a certain number of records to start the anonymization and publishing processes. The proposed scheme considers data stream DS as an ordered sequence of time-based sliding windows as follows:

$$DS = \{Sw_1, Sw_2, Sw_3, \dots, Sw_m\} \tag{27}$$

Where Sw_i represents the time of the current Sw under processing that exists for a specific period T and consists of a finite and varying number of records n. Poisson probability is employed as a prediction model to predict the data flow rate in the next sliding window, Sw_{i+1} based on the flow rate in the current Sw_i. The first sliding window buffer size Sw_i is set to an initial threshold value T bounded by a lower bound T_l and an upper bound T_u, where the k-anonymity model is applied to the data collected in Sw_i during T with TA, the processing time required to anonymize the data in Sw_i. As for the non-anonymized tuples, which belong to clusters with no sufficient records to meet the k-anonymity requirement, will be either incorporated into the

already anonymized clusters (reusable clusters) that can cover these records with the least information loss, or included in the subsequent sliding window Sw_{i+1} based on their expiry time TE as follows:

$$TE = Sw_i - TS - TA \tag{28}$$

Where TS is the time of storing the tuple in Sw_i, TE must be between the bounds for the current Sw $[T_l, T_u]$. It is continued to compute TE for each non-anonymized tuple, in addition to the expected arrival rate λ of the minimum number $(K - |AG|)$ of the similar records required to anonymize that group of non-anonymized tuples (AG) within the time interval of Sw_i using the lowest expiry time TE as follows:

$$\lambda = \frac{|AG_i|}{Sw_i} \cdot TE \tag{29}$$

This ensures avoidance of information loss caused by record expiry, since it does not anonymize the similar records if less than $(K - |AG|)$ tuples arrive. The arrival rate λ is then used to determine the probability of arrival of the needed N number of records, where $N = (K - |AG|)$ using the formula:

$$f(Sw_{i+1}, \lambda) = \Pr(j = 0 \dots N) = \frac{\lambda^i \cdot e^{-\lambda}}{j!} \tag{30}$$

Accordingly, the arrival probability of N or greater than N records in the stream within time TE is computed as:

$$1 - \sum_{i=0}^{N-1} Pr \tag{31}$$

The non-anonymized records will be included in the subsequent Sw and its size will be set to the lowest used expiry time if the arrival probability is greater than a pre-set probability threshold δ. Otherwise, they will be anonymized using a reusable cluster and the subsequent Sw size is set to a random time value or some initial threshold value within the time bound $[T_l, T_u]$. The proposed scheme handles the non-anonymized buffered tuples by either delaying them to the next buffering Sw or incorporating them into an anonymized cluster (reusable cluster) with similar privacy constraints, which reduces information loss to 1.95% in comparison to other solutions with an average information loss of 12.7% using the generic information loss metrics presented in equations (18), (19) and (20) [70]. The proposed model also applies the adaptive sliding window resizing that provides better time complexity. However, as it uses a clustering generalization-based method and k-anonymity with no restrictions on the SAs' values, the published data suffer from the attribute disclosure, similarity and sensitivity attacks, and valuable information loss. Besides, the approach was unable to effectively recover some of the suppressed tuples, because either their deadlines were exceeded. Moreover, the Sw size prediction for recovering those records was low, or the suitable reusable cluster could not be constructed before they expired. This is in addition to not considering the cases when anonymization might not be possible because no or few records exist in the stream.

V. PPTDP IN THE CLOUD ENVIRONMENTS

The incremental enormous amount of high-dimensional data generated from various sources generated a new data paradigm called Big Data, which has become a reality in the recent years at a wide range of fields. Big data aim to gather as many data as possible for data analysis and knowledge extraction purposes. This makes the privacy of the individuals, whose data are being collected and analyzed, is increasingly at risk [71]. The tremendous expansion of the cloud computing platforms provides a flexible infrastructure with powerful computation and storage resources, where users are enabled to handle such big data and their applications in a high scalable manner [72]. Studies have recently been directed towards adapting the previously-mentioned privacy preservation anonymization approaches for tabular data publishing in such emerging environments. Thus, these approaches are deployed in the cloud environments as they are without any modifications in order to handle data scalability and preserve data privacy. Authors in [73] employed a k-anonymity-based model to anonymize microdata before publishing them to the cloud service providers in order to be analyzed and mined. It anonymizes different QIDs for several cloud service providers depending on their variant background knowledge to ensure data anonymity [74].

In [75], a new iterative scalable k-anonymization model was introduced based on MapReduce paradigm [76] to preserve privacy of sharing individuals' data in the intercloud of Safer@Home welfare smart system [77] via k-anonymity. The proposed model used the distributed MapReduce paradigm and Mondrian multi-dimensional partitioning algorithm [30] to perform the anonymization process in a scalable way, while the data are managed in the Hadoop Distributed File System (HDFS). The model partitioned the data into ECs subsets, and then recursively partitioned each of these ECs into further subsets using the Mondrian algorithm until they satisfy the k-condition. A partitioned EC is said to satisfy the k-condition when its size is larger than or

equal to k, or smaller than or equal to $2k-1$. The other ECs are iteratively re-partitioned until the resultants satisfy the k-condition. Two MapReduce jobs carried out the anonymization process iteratively until the whole dataset is anonymized. SaCFRAPP was proposed in [78] as a scalable and cost-effective framework to preserve privacy of big data publishing on the cloud. The framework uses cloud-based MapReduce to perform the data anonymization process in high scalability and elasticity, where HDFS is used to manage the anonymous datasets before being released and published to other parties. SaCFRAPP consisted of four main modules; (1) Privacy Specification Interface (PSI) module, by which the data holders can specify the privacy requirements for anonymizing original datasets either by applying k-anonymity, l-diversity, or t-closeness; (2) Data Anonymization (DA) module, which utilizes MapReduce in anonymizing data using a generalization-based method; (3) Data Update (DU) module that anonymizes the data updates and adjusts the already anonymized datasets to ensure that the privacy preservation of the whole anonymized datasets is maintained. Finally, the Anonymous Data Sets Management (ADM) module, which retains anonymous datasets for data sharing, mining, and analytics rather than re-computing them repeatedly in order to save costs of both computation and storage resources that can be charged in the context of cloud computing.

Another scalable Two-Phase Top-Down Specialization (TPTDS) approach was proposed to anonymize large-scale datasets for privacy preservation using the MapReduce parallel data processing framework on cloud [79]. A group of MapReduce jobs was designed to perform the specialization computation for data anonymization in a scalable manner, which addresses the scalability problem of the Top-Down Specialization (TDS) approach [80]. TPTDS partitions the original datasets into smaller ones and anonymizes them in parallel, which generates intermediate outcomes. These intermediate outcomes are combined into one and then anonymized again to produce consistent anonymous datasets that satisfy the k-anonymity privacy principle. In [81], a scalable multi-dimensional anonymization approach based on MapReduce was introduced to preserve the privacy of big data over the cloud for scalability and cost-effectiveness. The proposed approach adopted the k-anonymity privacy model with the multi-dimensional scheme that considered multiple attributes together when generalizing domain values with the global recoding. The coefficient of variation [82] and the combination of the median of medians [83] with histogram were used in a scalable MapReduce-based algorithm to find the best splitting attribute and its suitable splitting point respectively for the data recursive partitioning. A group of MapReduce jobs was designed to perform the multi-dimensional anonymization on the partitioned datasets collaboratively in a recursive and highly scalable mechanism. The multi-dimensional anonymization approach partitioned the dataset into a set of non-overlapping multi-dimensional ECs and then generalized the QIDs of records. Consequently, the resultant anonymous dataset is said to satisfy the k-anonymity if the size of each EC is not less than k.

A hybrid scalable sub-tree anonymization approach over big data using MapReduce on cloud was proposed in [84], where the k-anonymity privacy model was adopted. The presented approach combined the Top-Down Specialization (TDS) [79-80] and Bottom-Up Generalization (BUG) [85] techniques to accomplish the sub-tree anonymization scheme, in which the entire child values of a non-leaf node or none in a domain hierarchy, were generalized to the node. The approach automatically determines either TDS or BUG will be used to anonymize the given dataset through a comparison between the "workload balancing point", which is a specific threshold K estimated from the evenly-distributed dataset, and a user-specified k-anonymity parameter k. If $k \geq K$, TDS is selected, otherwise BUG is selected, in order to minimize the required computations and achieve more efficient performance. A scalable MapReduce-based algorithm was then developed to perform either TDS or BUG in the big data anonymization process, exploiting the powerful computation abilities of the cloud.

Two enhancements were provided in [86] to the hybrid scalable sub-tree anonymization approach presented in [84]. The first one was to perform multiple generalization operations in one iteration round, by which the multiple generalization candidates can be considered in one round. The second was the skewness-aware workload balancing point adjustment, which handled the estimation of the workload balancing point K in case of the skewed data distribution, not only the even distribution as in [84]. The reason behind this was that the skewed data distribution can increase the size of ECs, leading to more violations of the k-anonymity requirement. A parameter $\gamma, 0 \leq \gamma \leq 1$ defined as the "adjustment factor", was used to adjust the value of K. Accordingly, the adjusted workload balancing point was equal to $\gamma \cdot K$, where the adjustment factor value γ depended on the data attributes distribution variance. These enhancements improved the parallelization, scalability and efficiency of BUG in the proposed approach. In [87], a research case study was conducted for a hospital management model to preserve privacy of health care big data in public cloud through data anonymization. The medical records were anonymized through a top-down specialization and button-up generalization hybrid anonymization algorithm in order to be transferred or published to other parties for further analysis or treatment research purposes.

VI. PPTDP IN IOT

Researchers in [88] have discussed the problem of finding a utility-aware privacy preservation solution in IoT applications from the practical and deployment perspective. A negotiation module was proposed within the IoT system, which aims to arriving at an agreement between the data producer and data consumer on the usage of the produced data containing sensitive users' data to provide their services in the IoT platform. By this agreement, the IoT platform creates a privacy preserving rule to enforce data privacy with utility. This rule consists of a set of SAs of the data to be published with their corresponding privacy preservation techniques. This includes hierarchical-based generalization, perturbation (addition of noise data) and randomization that can be applied using SafeMask, a proposed dynamic data masking solution [89]. SafeMask decodes all the data consumers'

requests to the IoT platform for user data, in which it reads the privacy rules and identifies the SAs with their corresponding privacy preservation technique. It then masks the sensitive data using the defined preservation technique in the rule. In [90], a new anonymization model was proposed to publish IoT data streams generated from multiple devices via partitioning under a specific delay constraint. The k-anonymity privacy model was enhanced to anonymize similar tuples in a cluster using partitioning under time-based sliding window and to anonymize tuples with missing values using representative values. The proposed model partitions the input tuples according to their attribute description, and then checks the size of the partition having expiring tuples. If this partition has enough tuples satisfying the k-anonymity requirement, the model either determines its new cluster generalization using the K-Nearest Neighbor (KNN) algorithm [91], or uses a reusable k-anonymous cluster over the same partition if exists, and then releases the tuples with this generalization. Otherwise, it merges the partitions based on their similarity, and creates cluster generalization for the resultant partition in order to publish the tuples. The Jaccard's similarity coefficient was used to measure the partitions similarity [92]. The information loss metrics presented in equations (18), (19) and (20) were used to measure the quality of the anonymization model.

J.L.H. Ramos, J.B. Bernabé and A.F. Skarmeta proposed another framework in [93] to enable privacy-aware data sharing on IoT environment. It provides a secure and privacy preserving framework for data sharing among smart objects in the IoT paradigm. A smart object may be represented by any entity that can generate (producer) and get (consumer) information. Each smart object is assumed to hold a partial identity in the data sharing context [94], anonymous credential systems [95] and Attribute-Based Encryption (ABE) [96]. In [97], an overview was provided about the privacy preserving techniques in IoT, where the privacy disclosure concerns in IoT were divided into two categories: data privacy and location privacy [98]. Data privacy presents the private information disclosure in the process of data acquisition, transmission, processing and publishing, whereas location privacy points to the location privacy of each node and the LBS in IoT environment. The author discussed how to reach the two kinds of privacy protection through the data anonymization models such as k-anonymity, l-diversity, and t-closeness. The data anonymization methods have preserved data privacy by satisfying the k-anonymity principle, and protected the location privacy information of individuals by using a trusted anonymous third party between the user and the LBS [99].

Another analytical study was presented in [100] for the privacy-preserving models and their use in the IoT infrastructure. The investigated privacy-preserving models were classified into four types; (1) general approaches for data privacy and k-anonymity, (2) homomorphic encryption, (3) group and ring signatures, and (4) Attribute Based Signatures (ABS) and ABE. The role of the basic approaches for data privacy and k-anonymity was discussed, in which data masking and hiding sensitive information were used for privacy protection. A solution for the combination of context aware access control and data transformation using k-anonymity to protect privacy was proposed in [101]. This context aware solution used k-anonymity to handle the identifiers of the records until each record cannot be distinguished within $k-1$ records, while the data publisher can transform the raw data using his privacy settings, i.e. masking or perturb the sensitive data. A privacy-preserving and security maintaining framework through the generation, collection, transfer, storage, processing and sharing of the sensor data from smart homes was presented in [102]. The proposed framework used the k-anonymity privacy model to achieve the privacy preservation level of the shared data using generalization and suppression. The framework consisted of three modules; data collector, data receiver and result provider, in addition to two storage units, which are the de-identified sensor data and the identifier dictionary storage units. Data collector module collects the sensors data and transfers them to the data cluster at regular intervals in a fast and secure manner using SSH transfer protocol [103]. Data receiver takes the input from the data collectors and performs an algorithmic function to classify the attributes of the data into primary/quasi-identifiers and non-identifier attributes according to an existing schema definition file. The primary/quasi-identifiers are hashed using SHA techniques [104-105], and stored together with their actual values in the identifier dictionary storage unit, if they do not already exist. The non-identifiers and the hashed primary/quasi-identifiers are stored in the de-identified storage unit. The result provider contains four sub-modules; the first is the access control sub-module, which provides system authentication and authorization to the end-user based on a set of rules, the second is the identifier retriever sub-module, which produces both actual and hashed values list of personal/quasi-identifiers requested by an authorized end-user by querying the identifier dictionary storage. The third sub-module is the transformer module that is responsible for ensuring the privacy of the shared data using the k-anonymity privacy model. The transformed dataset is said to satisfy k-anonymity if every combination of values in personally identifiable columns cannot be matched to fewer than k rows. The result processor is the fourth sub-module, which is responsible for starting a job on the de-identified storage and swapping the hashed personal/quasi-identifier values in the result set with the respective k-anonymized values, preserving the privacy of any shared data.

VII. DISCUSSION AND RESEARCH GAP

All the PPTDP models investigated in this study have not provided a discussion about the data analytical features of the anonymized data (i.e. outliers, granularity, cardinality, mean, variance, frequency, distance, etc.) nor how these analytical features were affected by the anonymization process. This deserves more consideration in the future as an interesting research direction in the field of PPTDP. However, the main focus of those models was to present a strong privacy preservation model for data publication, providing robust protection guarantees against the different privacy disclosure attacks in order to achieve safe and privacy-preserved publication of the individuals' data. The evaluation criterion was to measure the data utility of the

published data using different data quality and information loss metrics in order to retain as much data utilization capabilities as possible for any analysis or research purposes. Thus, all the data quality and information loss metrics applied by the models considered in this study were included, such as the Discernibility Metric (DM), the normalized average EC size metric (CAVG), the relative error, the distortion and the general information loss metrics. Additionally, the linking of multiple datasets and its associated potential for privacy risk disclosure violations have not been studied in the privacy preservation models analyzed in this study, which denotes another exciting research direction in PPTDP.

In the different studied privacy preservation models, the attributes' data type of the QIDs, either categorical or numerical, was not a concern, as they are processed by either the generalization or bucketization techniques. As for the SAs, the reflection of the attributes' data types was not explicitly discussed. This is because the restrictions introduced in the investigated privacy models in this study, like l-diversity and p-sensitive, can be applied on both data types. In addition, the SSA models have mostly conducted their experiments on a dataset having one categorical SA, although it can be deduced that their functionality can also be applied on a numerical SA. On the other hand, the experiments held for the studied MSA models included both categorical and numerical SAs. The exceptions were the FAANST and its extensions that were dedicated to work with numerical data only, t-closeness and (n, t)-closeness that were not prone to be used with numerical attributes as discussed earlier. This is due to difficulty in applying their property and the proximity breach occurred in such case [36]. Besides, the investigated SSA models have been practically tested and evaluated using SSA datasets only, in which their computations complexity has never been experimented using MSA datasets. This would make it difficult to provide a general conclusion about their applicability in MSA cases. However, it can be inferred that the functionality of some models can be applied on a MSA dataset, such as k-anonymity since no restriction is imposed on the sensitive values in each EC, and l-diversity that can apply its diversity condition on a MSA dataset. Others have functionality difficulties, making them inapplicable in MSA cases, like t-closeness and (n, t)-closeness.

On the other hand, most of the researches in preserving data stream privacy have been focusing on the accumulation count of tuples strategy with the generalization based-methods. Thus, these methods suffer from a number of shortcomings in terms of the information loss, real-time release of the published data, and weak data analysis capabilities and utilization. In addition, these researches do not restrict the published sensitive values with certain privacy requirements, which make the published tables vulnerable to many privacy-disclosure attacks. For example, consider Table IX(A) that shows an original microdata input stream, where the records represent the information of some patients in a medical dataset that need to be published. The attribute {Name} is an EI, attributes {Age, Gender, Zip code} are the QIDs and attribute {Disease} is the SA. Table IX(B) presents the resultant generalized table of Table IX(A), which satisfies 4-anonymity.

.

As shown in Table IX(B), the attribute {Name} is removed and the values of the QIDs {Age, Zip code} are generalized, producing three QI-groups with the following generalizations: <[20, 30], M, [15k, 25k]>, <[30, 45], F, [30k, 55k]>, and <[50, 60], F, [60k, 75k]>, including tuples 1-4, 5-8 and 9-12 respectively. Examples are: the age 25 and zip code 20k of tuple 2 have been replaced in the first QI-group by the intervals [20, 30] and [15k, 25k] respectively. The age 30 and zip code 35k of tuple 6 have been replaced in the second QI-group by the intervals [30, 45] and [30k, 55k] respectively. The age 55 and zip code 72k of tuple 11 have been replaced in the third QI-group by the intervals [50, 60] and [60k, 75k] respectively. Suppose that the attacker knows that Adam's age is 22 and lives in zip code 24k, such that his tuple is <22, M, 24k> from any publicly dataset or from his background knowledge and Table IX(A) is published. Through the linking between his knowledge and the available table, the attacker can determine that Adam's tuple is in EC1, which has hepatitis, flu, tonsillitis, and flu values for the disease SA. Thus, he can confidently conclude that Adam has flu with a high confidence ratio of 50% with the similarity attack. Table IX(C) determines the sensitivity levels of the sensitive values of the disease attribute in Table IX(A). As shown in Table IX(C), the sensitive values of the SA disease are categorized into three different sensitivity levels. The first sensitivity level included: stomach cancer, leukemia, esophagus cancer and HIV sensitive values, the second sensitivity level included: heart disease, hepatitis, and gastritis, and the third sensitivity level included: flu, and tonsillitis sensitive values. Furthermore, considering this table, the attacker can conclude that Adam has a disease that belongs to the 3rd sensitivity level with high confidence ratio of 75% with the sensitivity attack. Another instance, suppose that the attacker knows that Sandra's age is 42 and lives in zip code 54k, such that her tuple is <42, F, 54k>. Thereby, the attacker can determine that Sandra's tuple is included in EC2, which has gastritis, esophagus cancer, leukemia and stomach cancer values for the disease. Therefore, the attacker can confidently conclude that Sandra has a cancer disease with high confidence ratio of 75% with the similarity attack. Besides, the attacker can conclude that Sandra has a disease that belongs to the 1st sensitivity level with also high confidence ratio of 75% with the sensitivity attack.

The other researches that depend on bucketization-based methods suffer from membership and identity disclosure attacks, extra time overhead, and the breaking of correlation between the QIDs and SAs. The proposed adaptive buffer resizing approach that used time-based sliding window employed the k-anonymity privacy model, which has many deficiencies as explained

earlier. Thus, all of these work effectively with the nature of data as a stream, but they do not consider the privacy preserving restrictions and constraints on the published sensitive data. Therefore, more attention is required to balance the current trade-off situation between data utilization and privacy preservation in this field.

In addition, the discussed privacy preservation models in the cloud and IoT environments have adopted the k-anonymity model to ensure the privacy of the shared sensitive data. However, this decreases the published data utilization due to the resultant information and attributes correlation loss. Besides, this model has no restrictions on the published sensitive values, which makes the published data vulnerable to many privacy-disclosure attacks, specifically the attribute disclosure, similarity, skewness and sensitivity attacks. This opens the door for further required investigations from the research community to provide more efficient approaches that adopt more robust privacy-preserving models in such environments. Such required approaches are supposed to maintain data utility and cost-effectiveness and to achieve well-balancing between the scalability of large-scale datasets handling and sensitivity disclosure in order to provide a strong protection against the different privacy-disclosure attacks.

VIII. Conclusion

The Privacy-Preserving Tabular Data Publishing (PPTDP) has recently received a great attention in the research and applications due to its important role in data analysis, mining and decision-making purposes. In this paper, a comprehensive study is conducted in order to analyze and evaluate the different main data anonymization approaches that have been proposed in the PPTDP field, considering the SSA, MSA, and data streams publishing on the web, cloud and IoT environments. A detailed research gap with the possible future research directions has been discussed for the PPTDP field.

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] 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.

[3] N. Mohammed, R. Chen, B. Fung, and P. S. Yu, “Differentially private data release for data mining”, In Proc. ACM SIGKDD, pp. 493-501. ACM, 2011.

[4] C. Dwork, “Differential privacy”, In ICALP, 2006.

[5] C. Dwork, “Differential privacy: A survey of results”, In International Conference on Theory and Applications of Models of Computation, pp. 1-19, 2008.

[6] Y. Yang, Z. Zhang, G. Miklau, M. Winslett, and X. Xiao, “Differential privacy in data publication and analysis”, In Proc. ACM SIGMOD, pp. 601-606. ACM, 2012.

[7] M. Hardt, K. Ligett, and F. McSherry, “A simple and practical algorithm for differentially private data release”, In Proc. Advanced NIPS'12, pp. 2339-2347, 2012.

[8] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, “Differentially Private Data Publishing and Analysis: a Survey”, IEEE Transactions on Knowledge and Data Engineering, 2017.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis”, In Theory of Cryptography Conference, pp. 265-284, 2006.

[10] A. Blum, K. Ligett, and A. Roth, “A learning theory approach to noninteractive database privacy”, Journal of the ACM (JACM), 60(2), p.12, 2013.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] J.J. Kim and W.E. Winkler, “Masking microdata files”, In Proc. SRMS ASA, pp. 114–119, 1995.

[18] C.C Aggarwal, and P.S. Yu, “A condensation approach to privacy preserving data mining”. In Proc. ICEDT, pp. 183-199, 2004.

[19] R. Agrawal and R. Srikant, “Privacy-preserving data mining”, In ACM Sigmod Record, vol. 29, no. 2, pp. 439-450. ACM, 2000.

[20] J. Domingo-Ferrer, F. Sebé, and A. Solanas, “An anonymity model achievable via microaggregation”, In Workshop on Secure Data Management (pp. 209-218), 2008.

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

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

[23] 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.

[24] 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.
[25] 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.
[26] 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.
[27] T.M. Truta, and B. Vinay, "Privacy Protection: p-Sensitive k-Anonymity Property", In ICDE workshops, p. 94, 2006.
[28] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity", ACM TKDD, 1(1), p.3, 2007.
[29] R. J. Bayardo, and R. Agrawal, "Data privacy through optimal k-anonymization", In Proc. ICDE 2005, (pp. 217-228). IEEE, 2005.
[30] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity", In Proc. ICDE'06, (pp. 25-25). IEEE, 2006.
[31] X. Sun, L. Sun, H. Wang, "Extended k-anonymity models against sensitive attribute disclosure", Computer Communications 34, pp.526–535, 2011.
[32] N. Li, T. Li, S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity", In Proc. ICDE, pp. 106-115. IEEE, 2007.
[33] 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.
[34] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A new privacy measure for data publishing", IEEE TKDE, 22(7), pp.943-956, 2010.
[35] F. Bacchus, A. Grove, D. Koller, and J.Y. Halpern, "From statistics to beliefs", In AAAI, pp. 602-608. 1992.
[36] 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.
[37] 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.
[38] X. Xiao, Y. Tao, "Anatomy: simple and effective privacy preservation", In Proc. VLDB, pp.139–150, 2006.
[39] 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.
[40] 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.
[41] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A.W.C. Fu, "Utility-based anonymization using local recoding", In Proc. 12th ACM SIGKDD, pp. 785–790.ACM, 2006.
[42] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing location-based identity inference in anonymous spatial queries", IEEE TKDE 19(12), pp.1719-1733, 2007.
[43] M.F. Mokbel, C.W. Chow, W.G. Aref, "The new Casper: query processing for location services without compromising privacy", In Proc. VLDB, pp.763–774, 2006.
[44] 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.
[45] H. Wang, "Privacy-preserving data sharing in cloud computing", Journal of Computer Science and Technology, 25(3), pp.401-414, 2010.
[46] V. Rastogi, D. Suciu and S. Hong, "The boundary between privacy and utility in data publishing", In Proc. VLDB, pp.531-542, 2007.
[47] M. Rajaei and M.S. Haghjoo, "An improved Ambiguity+ anonymization technique with enhanced data utility", In Proc. IKT, pp. 1-7. IEEE, 2015.
[48] T. Li, N. Li, J. Zhang and I. Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE TKDE, 24(3), pp.561-574, 2012.
[49] S. Kiruthika, Dr. M. Mohamed Rasee, "Enhanced Slicing Models for Preserving Privacy in Data Publication", ICCTET, pp. 406-409. IEEE, 2013.
[50] F. Luo, J. Han, J. Lu and H. Peng, "ANGELMS: A Privacy Preserving Data Publishing Framework for Microdata with Multiple Sensitive Attributes", ICIST, pp. 393-398. IEEE, 2013.
[51] 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.
[52] C.N. Sowmyarani, G.N. Srinivasan, "A Robust Privacy Preserving Model for Data Publishing", ICCCI, pp. 1-6. IEEE, 2015.
[53] Y. Wu, X. Ruan, S. Liao, and X. Wang, "P-cover k-anonymity model for protecting multiple sensitive attributes", In Proc. ICCSE, pp. 179-183. IEEE, 2010.
[54] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity", In Proc. ACM SIGMOD, pp. 49-60. ACM, 2005.
[55] J.W Byun, Y. Sohn, E. Bertino, and N. Li, "Secure anonymization for incremental datasets", In Workshop on Secure Data Management, pp. 48-63, 2006.
[56] Y. Ye, Y. Liu, D. Lv, and J. Feng, "Decomposition: Privacy Preservation for Multiple Sensitive Attributes", In Proc. DASFAA, pp. 486-490, 2009.
[57] J. Li, B.C. Ooi, and W. Wang, "Anonymizing streaming data for privacy protection", in Proc. ICDE, pp.1367-1369, 2008.
[58] J. Cao, B. Carminati, E. Ferrari, and K.L. Tan, "CASTLE: A Delay-Constrained Scheme for Ks-Anonymizing Data Streams", in Proc. ICDE, pp. 1376-1378. IEEE, 2008.
[59] P. Wang, J. Lu, L. Zhao and J. Yang, "B-CASTLE: An Efficient Publishing Algorithm for K-anonymizing Data Streams", Second WRI GCIS vol. 2, pp. 132-136. IEEE, 2010.

[60] G. Yang, J. Yang, J. Zhang, and Y. Chu, "Research on data streams publishing of privacy preserving", In ICITIS, pp. 199-202. IEEE, 2010.

[61] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "A framework for clustering evolving data streams", In Proc. VLDB, pp. 81-92, 2003.

[62] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases", In ACM Sigmod Record, vol. 25, no. 2, pp. 103-114. ACM, 1996.

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

[64] P. Wang, L. Zhao, J. Lu, and J. Yang, "SANATOMY: Privacy preserving publishing of data streams via anatomy." In ISIP, pp. 54-57. IEEE, 2010.

[65] 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.

[66] W. Wang, J. Li, C. Ai, and Y. Li, "Privacy protection on sliding window of data streams." In Proc. CCNAW, CollaborateCom, pp. 213-221. IEEE, 2007.

[67] H. Zakerzadeh, and S.L. Osborn, "FAANST: fast anonymizing algorithm for numerical streaming data." In DPMASS, Springer Berlin Heidelberg, pp. 36-50., 2011.

[68] H. Zakerzadeh and S.L. Osborn. "Delay-sensitive approaches for anonymizing numerical streaming data." International journal of information security 12(5), pp.423-437, 2013.

[69] E. Mohammadian, M. Noferesti, and R. Jalili, "FAST: Fast Anonymization of Big Data Streams." In Proc. ICBDSC, p. 23. ACM, 2014.

[70] A.B. Sakpere, and A.V. Kayem, "Adaptive buffer resizing for efficient anonymization of streaming data with minimal information loss", In Proc. ICISSP, pp.1-11. SCITEPRESS, 2015.

[71] J. Soria-Comas, and J. Domingo-Ferrer, "Big data privacy: challenges to privacy principles and model", Data Science and Engineering, 1(1), pp.21-28, 2016.

[72] S. Vennila, and J. Priyadarshini, "Scalable Privacy Preservation in Big Data a Survey", Procedia Computer Science, 50, pp.369-373, 2015.

[73] J. Wang, Y. Zhao, S. Jiang, and J. Le, "Providing privacy preserving in cloud computing", In Proc. ICHSI, pp. 472-475. IEEE, 2010.

[74] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables", In Proc. IEEE ICDE, pp. 116-125. IEEE, 2007.

[75] A. Chakravorty, T. W. Wlodarczyk, and C. Rong, "A scalable k-anonymization solution for preserving privacy in an aging-in-place welfare intercloud", In Proc. IC2E, pp. 424-431. IEEE, 2014.

[76] J. Dean, and S. Ghemawat, "MapReduce: simplified data processing on large clusters" Communications of the ACM, 51(1), pp.107-113, 2008.

[77] T. W. Wlodarczyk, C. Rong, and D. Waage, "Challenges in healthcare and welfare intercloud", In Proc. IDCTA, pp. 45-48. IEEE, 2011.

[78] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen, "SaC-FRAPP: a scalable and cost-effective framework for privacy preservation over big data on cloud". Concurrency and Computation: Practice and Experience, 25(18), pp.2561-2576, 2013.

[79] X. Zhang, L.T. Yang, C. Liu, and J. Chen, "A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud", IEEE Transactions on Parallel and Distributed Systems, 25(2), pp.363-373, 2013.

[80] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Engineering, 19(5), pp. 711-725, 2007.

[81] X. Zhang, C. Yang, S. Nepal, C. Liu, W. Dou, and J. Chen, "A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud", In Proc. CGC, pp. 105-112. IEEE, 2013.

[82] W.A. Hendricks and K.W. Robey, "The sampling distribution of the coefficient of variation" The Annals of Mathematical Statistics, vol. 7, no. 3, pp. 129-132, 1936.

[83] M. Blum, R.W. Floyd, V. Pratt, R. L. Rivest and R. E. Tarjan, "Time bounds for selection", Journal of Computer and System Sciences, vol.7, no. 4, pp. 448-461, 1973.

[84] X. Zhang, C. Yang, S. Nepal, C. Liu, W. Dou, and J. Chen, "Combining top-down and bottom-up: scalable sub-tree anonymization over big data using MapReduce on cloud", In Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE International Conference on, pp. 501-508. IEEE, 2013. (Same authors of 6 and 66).

[85] W. Ke, P.S. Yu and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection," In Proc. ICDM'04, pp. 249-256. IEEE, 2004.

[86] X. Zhang, C. Yang, S. Nepal, C. Liu, W. Dou, and J. Chen, "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud", Journal of Computer and System Sciences, 80(5), pp.1008-1020, 2014.

[87] P. S. Rose, J. Visumathi, and H. Haripriya, "Research Paper on Privacy Preservation by Data Anonymization in Public Cloud for Hospital Management on Big data", International Journal of Advanced Computer Technology (IJACT), 2016.

[88] A. Ukil, S. Bandyopadhyay, J. Joseph, V. Banahatti, and S. Lodha, "Negotiation-based privacy preservation scheme in internet of things platform", In Proc. ICSIOT, pp. 75-84. ACM, 2012.

[89] V. Banahatti, S. Lodha, "SafeMask", In Proc. TACTICS, 2009.

[90] A. Otgonbayar, Z. Pervez, and K. Dahal, "Toward Anonymizing IoT Data Streams via Partitioning", In Proc. MASS, pp. 331-336. IEEE, 2016.

[91] L. E. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p.1883, 2009.

[92] J. Leskovec, A. Rajaraman, and J. D. Ullman, "Mining of massive datasets", Cambridge University Press, 2014.

[93] J.L.H. Ramos, J.B. Bernabé, and A.F. Skarmeta, "Towards Privacy-Preserving Data Sharing in Smart Environments", In Proc. IMIS, pp. 334-339. IEEE, 2014.

[94] M. Hansen, P. Berlich, J. Camenisch, S. Clauß, A. Pfitzmann, and M. Waidner, "Privacy-enhancing identity management", Information security technical report, 9(1), pp.35-44, 2004.

[95] J. Camenisch and A. Lysyanskaya, "An efficient system for non-transferable anonymous credentials with optional anonymity revocation", In Proc. ICTACT, pp. 93–118, 2001.
[96] A. Sahai and B. Waters, "Fuzzy identity-based encryption", In Proc. ICTACT, pp. 457–473, 2005.
[97] P. S. Wang, "Privacy Preserving Techniques in the Internet of Things", In Applied Mechanics and Materials, Vol. 427, pp. 2466-2469. Trans Tech Publications, 2013.
[98] N.H. Li, N. Zhang, S.K. Das, B. Thuraisingham, "Privacy preservation in wireless sensor networks: A state-of-the-art survey", Ad hoc Networks, 7(8), pp. 1501-1514, 2009.
[99] C.Y. Chow, M.F. Mokbel and W.G. Aref, "Casper*: Query processing for location service without compromising privacy", ACM Transactions on Database Systems, 34(4), pp. 1-24, 2009.
[100] L. Malina, J. Hajny, R. Fujdiak, and J. Hosek, "On perspective of security and privacy-preserving solutions in the internet of things", Computer Networks, 102, pp.83-95, 2016.
[101] X. Huang, R. Fu, B. Chen, T. Zhang, A. Roscoe, "User interactive internet of things privacy preserved access control", In Proc. IEEE ITST, pp. 597–602. IEEE, 2012.
[102] A. Chakravorty, T. Wlodarczyk, and C. Rong, "Privacy Preserving Data Analytics for Smart Homes", In Security and Privacy Workshops (SPW), pp. 23-27. IEEE, 2013.
[103] T. Yloenen, "SSH - Secure Login Connections over the Internet", 6th USENIX UNIX Security Symposium, 1996.
[104] F. Mendel, N. Pramstaller, C. Rechberger, and V. Rijmen, "Analysis of step-reduced SHA-256", In International Workshop on Fast Software Encryption, pp. 126-143, 2006.
[105] H. Gilbert, and H. Handschuh, "Security Analysis of SHA-256 and Sisters", In Selected areas in Cryptography, vol. 3006, pp.175-193, 2004.

Fig. 1. An abstract architecture of Privacy-Preserving Tabular Data Publishing (PPTDP)
Fig. 2. The categorization of PPDP approaches for tabular data (PPTDP)

TABLE I. ORIGINAL RAW DATA FORMAT IN PPTDP.

Name	Age	Gender	Zip code	Disease
Alex	27	M	16k	hepatitis
James	25	M	18k	flu
Adam	22	M	24k	flu
Maria	34	F	38k	gastritis
Sandra	42	F	54k	leukemia
Andy	38	F	40k	stomach cancer
Martina	52	F	68k	leukemia
Victoria	57	F	62k	heart disease
Catty	55	F	72k	HIV

TABLE II. EXAMPLE FOR 3-ANONYMITY TABLE OF TABLE I.

Age	Gender	Zip code	Disease
[20, 30]	M	[16k, 25k]	hepatitis
[20, 30]	M	[16k, 25k]	flu
[20, 30]	M	[16k, 25k]	flu
[30, 45]	F	[30k, 55k]	gastritis
[30, 45]	F	[30k, 55k]	leukemia
[30, 45]	F	[30k, 55k]	stomach cancer
[50, 60]	F	[60k, 75k]	leukemia
[50, 60]	F	[60k, 75k]	heart disease
[50, 60]	F	[60k, 75k]	HIV

TABLE III. EXAMPLE FOR 2-DIVERSITY TABLE OF TABLE I.

Age	Gender	Zip code	Disease
[20, 30]	M	[16k, 25k]	hepatitis
[20, 30]	M	[16k, 25k]	flu
[20, 30]	M	[16k, 25k]	flu
[30, 60]	F	[30k, 75k]	gastritis
[30, 60]	F	[30k, 75k]	leukemia
[30, 60]	F	[30k, 75k]	stomach cancer
[30, 60]	F	[30k, 75k]	leukemia
[30, 60]	F	[30k, 75k]	heart disease
[30, 60]	F	[30k, 75k]	HIV

TABLE IV (A). THE QUASI-IDENTIFIER TABLE (QIT) OF THE ANONYMIZED TABLES OF TABLE I.

Age	Gender	Zip code	Group-ID
27	M	16k	1
25	M	18k	1
22	M	24k	1
34	F	38k	2
42	F	54k	2
38	F	40k	2
52	F	68k	3
57	F	62k	3
55	F	72k	3

TABLE IV (B). THE SENSITIVE TABLE (ST) WITH 2-DIVERSITY OF THE ANONYMIZED TABLES OF TABLE I.

Group-ID	Disease	Count
1	hepatitis	1
1	flu	2
2	gastritis	1
2	leukemia	1
2	stomach cancer	1
3	leukemia	1
3	heart disease	1
3	HIV	1

TABLE V. COMPARISON BETWEEN SSA PRIVACY MODELS WITH RESPECT TO PRIVACY ATTACKS

Privacy Models	MD	ID	AD	SiA	SkA	SeA
k-anonymity	No	No	Yes	Yes	Yes	Yes
p-Sensitive k-Anonymity	No	No	If p=1 or 2	Yes	Yes	Yes
l-diversity	No	No	If l =1 or 2	Yes	Yes	Yes
p+ sensitive k-anonymity	No	No	No	No	Yes	Yes
(p, α) sensitive k-anonymity	No	No	No	No	No	Yes
t-closeness	No	No	No	No	No	Yes
(n, t)-closeness	No	No	No	No	No	Yes
(w, γ, k)-anonymity	No	No	No	No	No	No
Anatomy	Yes	Yes	No	Yes	Yes	Yes
Permutation Anonymization (PA)	No	No	No	Yes	Yes	Yes
De-clustering	Yes	Yes	No	Yes	No	Yes
Ambiguity, PriView and Ambiguity+	No	No	No	Yes	Yes	Yes

MD= Membership Disclosure, ID= Identity Disclosure, AD= Attribute Disclosure, SiA= Similarity Attack, SkA= Skewness Attack, SeA= Sensitivity Attack.

TABLE VI. GENERAL COMPARISON BETWEEN SSA PRIVACY MODELS

Privacy Models	QID Processing	SAs Restriction	Privacy Guarantees	Advantages	Additional Issues / Challenges
k-anonymity	Generalization	None	Generalizing QIDs	Prevents MD and ID	Curse of dimensionality, huge info. loss and QIDs correlation loss
p-Sensitive k-Anonymity	Generalization	p-Sensitive	Generalizing QIDs, imposing p-Sensitive restriction on SAs	Prevents MD, ID and AD	Curse of dimensionality, huge info. loss and QIDs correlation loss
l-diversity	Generalization	l-diversity	Generalizing QIDs, imposing new l-diversity restriction on SAs	Prevents MD, ID and AD, increases the diversity of the SA's values in each EC.	Curse of dimensionality, huge info. loss, QIDs correlation loss and difficulty to achieve high l values.
p+ sensitive k-anonymity	Generalization	p distinct categories	Generalizing QIDs, imposing p distinct categories restriction on SAs	Prevents MD, ID, AD and SiA	High info. loss, QIDs correlation loss and did not avoid SkA and SeA
(p, α) sensitive k-anonymity	Generalization	p distinct SA values with their total weight is at least α in each EC	Generalizing QIDs, imposing p distinct SA values with their total weight is at least α	Prevents MD, ID, AD, SiA and SkA	High info. loss, QIDs correlation loss and still faces SeA
t-closeness	Generalization	t-closeness distribution	Defining a semantic distribution distance for SAs in each EC	Prevents MD, ID, AD, SiA and SkA	Faces SeA, inappropriate with data tables having many numerical attributes, difficulty to define a procedure to enforce t-closeness and greatly degrades the data utility
(n, t)-closeness	Generalization	(n, t)-closeness distribution	Defining a semantic distribution distance for SAs in each two natural superset ECs	Prevents MD, ID, AD, SiA and SkA and provides better data utility than t-closeness	Still causes proximity breach in a table of numeric SAs and faces SeA
(w, γ, k)-anonymity	Generalization	Each EC contains SA values, their average weight is at least w, with its similarity is at most γ	Generalizing QIDs, restricting the weight and similarity of SA values	Prevents MD, ID, AD, SiA and SkA and SeA, applied for both numeric and categorical SAs	Considerable info. loss, QIDs correlation loss and long computations time
Anatomy	None	l-diversity	The division of the QIT and ST, imposing l-diversity restriction on SAs	Allows more effective data accuracy and utilization	Faces MD and ID, loses QIDs & SAs correlation and when the number of the recurring sensitive value is so huge that decreases the number of distinct sensitive values in each EC
Permutation Anonymization (PA)	Random Permutation	l-diversity	The division of the QIT and ST, imposing l diversity restriction on SAs and the QIDs random permutation	Prevents MD and ID, allows more effective data accuracy and utilization	Loses QIDs & SAs correlation and the applied restriction did not avoid SiA, SkA and SeA
De-clustering	None	Maximizes the number of distinct sensitive values using dissimilarity fn.	EC contains different number of records and max number of distinct sensitive values	Prevents AD and maintains better data accuracy and utilization	Faces MD and ID, used a restriction that did not avoid SiA, SkA and SeA
Ambiguity	Releases each QID in a separated table	l-diversity	The division between all QIDs and SAs and imposing l-diversity restriction on SA	Prevents MD, ID and AD and provides better data utility	Loses correlation between all QIDs, between QIDs & SAs and faces SiA, SkA and SeA
PriView	Releases QIDs in two tables	l-diversity	The division between some of QIDs and SAs and imposing l-diversity restriction on SA	Prevents MD, ID and AD and provides more better data utility than Ambiguity	Loses correlation between some QIDs and between QIDs & SAs and faces SiA, SkA and SeA
Ambiguity+	Releases QIDs in two tables	l-diversity	The division between some of QIDs and SAs and imposing l-diversity restriction on SA	Prevents MD, ID and AD and provides more better data utility than Ambiguity	Loses correlation between some QIDs and between QIDs & SAs and faces SiA, SkA and SeA

MD= Membership Disclosure, ID= Identity Disclosure, AD= Attribute Disclosure, SiA= Similarity Attack, SkA= Skewness Attack, SeA= Sensitivity Attack.

TABLE VII. COMPARISON BETWEEN MSA PRIVACY MODELS WITH RESPECT TO PRIVACY ATTACKS.

Privacy Models	MD	ID	AD	SiA	SkA	SeA
Slicing, Mondrian Slicing and Suppression Slicing	No	No	No	Yes	Yes	Yes
ANGELMS	No	No	No	Yes	Yes	Yes
Additive Noise Approach	No	No	No	Yes	Yes	Yes
(p^+)-sensitive t-closeness	No	No	No	No	No	No
P-cover k-anonymity	No	No	No	Yes	Yes	Yes
Decomposition	Yes	Yes	No	No	Yes	Yes

MD= Membership Disclosure, ID= Identity Disclosure, AD= Attribute Disclosure, SiA= Similarity Attack, SkA= Skewness Attack, SeA= Sensitivity Attack.

TABLE VIII. GENERAL COMPARISON BETWEEN MSA PRIVACY MODELS.

Privacy Models	QID Processing	SAs Restriction	Privacy Guarantees	Advantages	Additional Issues / Challenges
Slicing	Generalization and random permutation	l-diversity and random permutation	Partitioning the data both horizontally and vertically, generalizing QIDs, imposing l-diversity restriction on SAs, and random permutation	Prevents MD, ID and AD, handles high-dimensional data, preserves better data utility, and maintains correlation between QIDs and SAs	The cases of having different tuples have the same QIDs and SAs values and give the same original tuple after the random permutation within EC, and still faces SiA, SkA and SeA
Mondrian and Suppression Slicing	Generalization and random permutation	l-diversity and random permutation	Partitioning the data both horizontally and vertically, generalizing QIDs, imposing l-diversity restriction on SAs, and random permutation	Prevents MD, ID and AD, provides better data utility than Slicing, handles high-dimensional data and maintains correlations between QIDs and SAs	Faces SiA, SkA and SeA
ANGELMS	Generalization	Partitioned into several SAs tables satisfying l-diversity	Partitioning attributes into several SAs tables and one QIDs table, generalizing QIDs, imposing l-diversity restriction on SAs	Prevents MD, ID and AD	Loses QIDs & SAs correlation and considerable info. loss
Additive Noise Approach	Generalization	Adding $(l-1)$ diverse random selected noise value to the actual value.	Generalizing QIDs, satisfying the l-diversity restriction on the value set of each SA	Prevents MD, ID, and AD, increases the diversity of the SA's values in each tuple and maintains more attributes correlation	Faces SkA, SiA and SeA and decreases data utility with high l values
(p^+)-sensitive t-closeness	Generalization	t-closeness distribution on the distinct sensitivity level of each SA.	Generalizing QIDs and distributing the sensitivity level of each SA, such that each EC has at least p distinct sensitivity levels under the defined threshold t	Combines the advantages of the t-closeness and p-sensitive k-anonymity approaches and prevents MD, ID, AD, SiA, SkA and SeA	Considerable info. loss from QIDs generalization
P-cover k-anonymity	Generalization	MSA-P-diversity principle	Generalizing QIDs, satisfying the MSA-P-diversity restriction among MSA	Prevents MD, ID and AD	High info. loss and faces SiA, SkA and SeA
Decomposition	None	MSA diversity $((l_1, l_2, \dots, l_d)$-diversity) principle	Each SA-group contains at least l distinct sensitive values, tuples within the same group will share the union of their sensitive values and maximize the number of such SA-groups as much as possible	Prevents AD and SiA, maintains attributes correlation and allows more effective data utilization	Faces MD, ID, SkA and SeA

MD= Membership Disclosure, ID= Identity Disclosure, AD= Attribute Disclosure, SiA= Similarity Attack, SkA= Skewness Attack, SeA= Sensitivity Attack.

TABLE IX (A). ORIGINAL MICRODATA INPUT STREAM.

Name	Age	Gender	Zip code	Disease
Alex	27	M	15k	hepatitis
James	25	M	20k	flu
Tony	20	M	22k	tonsillitis
Adam	22	M	24k	flu
Maria	34	F	30k	gastritis
Sophie	30	F	35k	esophagus cancer
Sandra	42	F	54k	leukemia
Andy	38	F	40k	stomach cancer
Martina	50	F	60k	leukemia
Victoria	57	F	64k	heart disease
Catty	55	F	72k	HIV
Emilia	60	F	70k	heart disease
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

TABLE IX (B). 4-ANONYMITY GENERALIZED TABLE OF TABLE IX (A).

Age	Gender	Zip code	Disease
[20, 30]	M	[15k, 25k]	hepatitis
[20, 30]	M	[15k, 25k]	flu
[20, 30]	M	[15k, 25k]	flu
[20, 30]	M	[15k, 25k]	tonsillitis
[30, 45]	F	[30k, 55k]	gastritis
[30, 45]	F	[30k, 55k]	esophagus cancer
[30, 45]	F	[30k, 55k]	leukemia
[30, 45]	F	[30k, 55k]	stomach cancer
[50, 60]	F	[60k, 75k]	leukemia
[50, 60]	F	[60k, 75k]	heart disease
[50, 60]	F	[60k, 75k]	HIV
[50, 60]	F	[60k, 75k]	heart disease
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

TABLE IX (C). THE SENSITIVITY LEVELS TABLE OF ATTRIBUTE DISEASE.

Sensitivity Level	Sensitive Values
1	stomach cancer, leukemia, esophagus cancer, HIV.
2	heart disease, hepatitis, gastritis.
3	flu, tonsillitis.

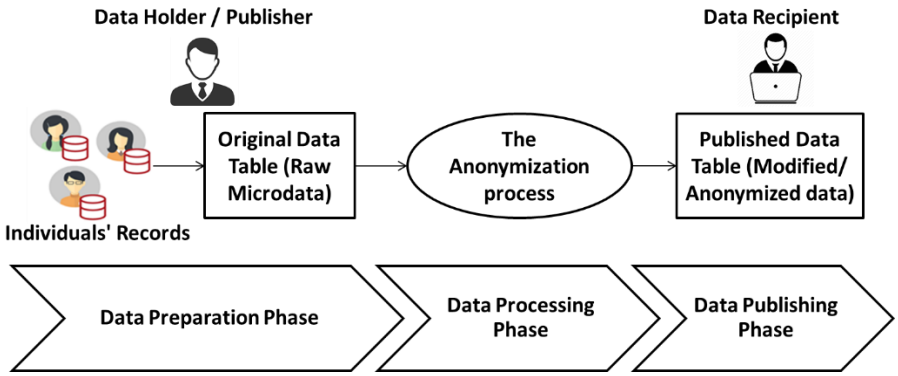

Fig. 1.

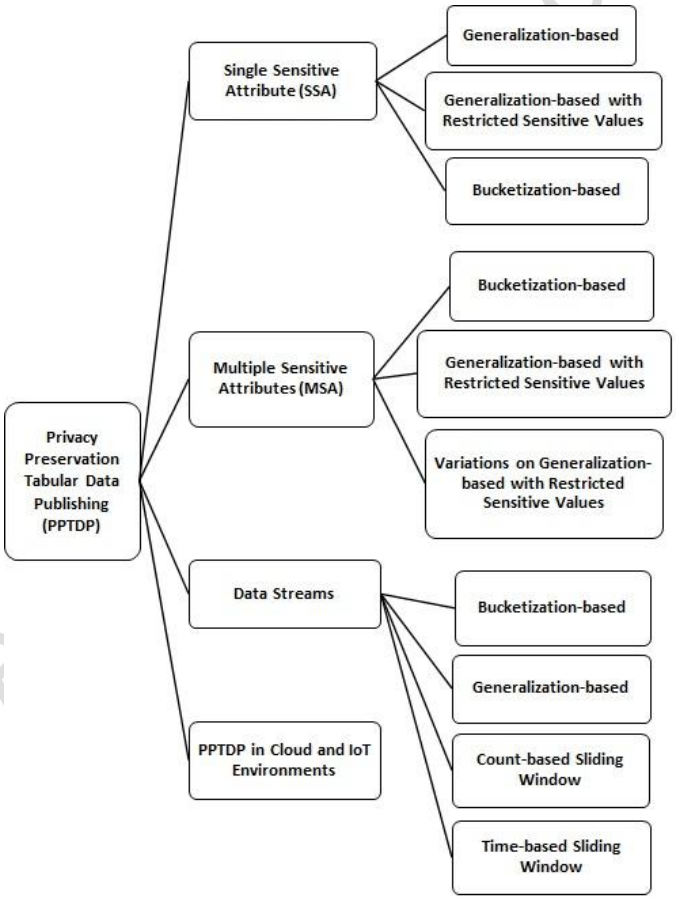

Fig. 2..