Systematic Evaluation of Probabilistic k-Anonymity for Privacy Preserving Micro-data Publishing and Analysis

Navoda Senavirathne[1] and Vicenç Torra[2]
[1]*School of Informatics, University of Skövde, Sweden*
[2]*Department of Computer Science, University of Umeå, Sweden*

Keywords: Data Privacy, Anonymization, Statistical Disclosure Control, Privacy Preserving Machine Learning.

Abstract: In the light of stringent privacy laws, data anonymization not only supports privacy preserving data publication (PPDP) but also improves the flexibility of micro-data analysis. Machine learning (ML) is widely used for personal data analysis in the present day thus, it is paramount to understand how to effectively use data anonymization in the ML context. In this work, we introduce an anonymization framework based on the notion of "probabilistic k-anonymity" that can be applied with respect to mixed datasets while addressing the challenges brought forward by the existing syntactic privacy models in the context of ML. Through systematic empirical evaluation, we show that the proposed approach can effectively limit the disclosure risk in micro-data publishing while maintaining a high utility for the ML models induced from the anonymized data.

1 INTRODUCTION

Data anonymization facilitates Privacy Preserving Data Publishing (PPDP) which allows the data controllers to share the data publicly or with specific third parties with minimal privacy implications. In data anonymization, the underlying micro-data (personal data) are irrecoverably transformed so that the risk of re-identifying the individuals (identity disclosure) is minimized along with the risk of inferring their sensitive characteristics (attribute disclosure). Data anonymization not only supports PPDP but it also introduces more flexibility into micro-data processing and analysis in the light of new privacy laws. According to recital 26 of the General Data Protection Regulation (GDPR), the data protection principles do not apply to anonymized data. To achieve anonymization under GDPR, re-identification (singling out) of an individual must be impossible under all the means reasonably likely to be used either by the data controller or by any other party directly or indirectly. Once the data are anonymized, they are allowed to be freely used, shared and monetized without the usual restrictions apply to the raw micro-data. Hence, by anonymizing the micro-data an organization can earn numerous advantages such as a) disclosure risks minimization, b) avoiding GDPR compliance risks and, c) improving the flexibility of data publishing and analysis. Due to these advantages anonymization seems to be a sensible approach for handling micro-data. Anonymized data are then subject to further analysis in order to facilitate reporting, knowledge extraction and/ or decision making.

In the present day, ML is employed in a wide variety of domains where micro-data are used for training the ML models. Moreover, in recent years there has been numerous research highlighting the privacy vulnerabilities of ML models trained on micro-data such as membership inference, attribute inference etc. (Al-Rubaie and Chang, 2019). As a mitigating strategy for privacy leakage via ML models, differential privacy (DP) based ML algorithms are proposed in the literature (Ji et al., 2014). They are aimed at limiting the effect of training data on the final ML model via injecting noise into the training process thus minimizing leakage of sensitive information. Even though DP is widely used to achieve privacy preserving ML (PPML), it has a significant drawback apart from the conventional challenges of implementing DP; such as utility loss due to noise injection, the complexity of estimating function sensitivity or ambiguity in deciding on privacy parameter (ε). That is differentially private ML models are assumed to be trained on the original micro-data thus requiring the data controllers to adhere to the data protection principles introduced in GDPR which limit the flexibility of further processing of micro-data (i.e., limitation on data retention and purpose of processing, the requirement for data in-

Senavirathne, N. and Torra, V.
Systematic Evaluation of Probabilistic k-Anonymity for Privacy Preserving Micro-data Publishing and Analysis.
DOI: 10.5220/0010560703070320
In *Proceedings of the 18th International Conference on Security and Cryptography (SECRYPT 2021)*, pages 307-320
ISBN: 978-989-758-524-1

tegrity and accuracy followed by transparency in data collection and accountability). On the other hand, if the underlying training data are already anonymized they are no longer considered as personal data thus providing data controllers and analysts more flexibility. Motivated by the above discussed advantages of data anonymization, it is plausible that the data controllers anonymize the data before publishing them for any data analysis task.

The field of statistical disclosure control (SDC) has introduced different privacy models and anonymization algorithms to support PPDP. k-Anonymity (Samarati, 2001) is one of the most widely used syntactic privacy models which ensures record indistinguishability within a set of k records as far as their quasi-identifier attributes (QIDs) are concerned thus minimizing the risk of re-identification. QIDs are a subset of attributes that can be used to re-identify the individuals in a published dataset uniquely, e.g., it is shown that simple demographic attributes such as birthday, zip code and, gender can be used together to re-identify about 87% of the US population (Sweeney, 2000). k-Anonymity relies on reducing the granularity of the QIDs to satisfy the indistinguishability requirement. Data generalization and suppression techniques are used for categorical QIDs whereas, microaggregation or rounding are used for numerical attributes to achieve k-anonymity (Sweeney, 2002) (Domingo-Ferrer and Mateo-Sanz, 2002). Moreover, concerning categorical data generalization, it is always not possible to obtain semantically meaningful categories for all the QIDs. Nevertheless, most of the real-world datasets contain both numerical and categorical QIDs (i.e., mixed datasets) thus necessitating us to use a combination of the above techniques for better data utility (e.g, use microaggregation on numerical data and generalization on categorical data) which makes the anonymization process tedious. According to Aggarwal (Aggarwal, 2005) when the number of QIDs is large, most of the attribute values have to be suppressed in order to satisfy k-anonymity conditions which degrade the data utility significantly. Utility loss of data caused by anonymization negatively impacts any analysis done on the anonymized data thereafter.

Since the k-Anonymity based privacy models are mainly focused on PPDP, they bring up unforeseen challenges when used in the context of ML where the ML models are induced from anonymized training data. This is in addition to the utility loss encountered by the ML models trained on the anonymized data. For example, k-Anonymity based privacy models make structural changes to QID attributes by the means of data suppression and generalization when it is applied to the categorical data. This brings up the below mentioned practical challenges as explained by (Senavirathne and Torra, 2020).

- Data insufficiency due to suppression - Due to data suppression a given ML algorithm might not have enough data to learn a meaningful pattern. Therefore, we may need to employ data imputation at the data preprocessing phase of the ML pipeline to approximate these suppressed values. This could intensify the accuracy loss of the ML models trained on anonymized data.
- Previously unseen attributes values due to generalization - Feature vectors pass into the deployed ML models must have the same attribute domains as the training data. However, when data generalization is used, it changes the attribute domains by replacing the existing values with new values from the top of the generalization hierarchy. Hence, anonymized training data are deprived of some attribute values that exist in the general population. At the inference phase of the ML pipeline, if a user submits a feature vector containing previously unseen attribute values, the ML model will fail to process that input. To address this issue the ML model owners have to use an API to transform user's raw data into the generalized format required by the ML model.

Addressing the above mentioned challenges require additional efforts in the ML process without any guarantee on model utility improvement. It is also shown that despite the distortion introduced to the underlying data, k-anonymity still suffers from privacy vulnerabilities. Thus some enhancements are proposed to k-anonymity. Out of them the most widely known methods are l-diversity (Machanavajjhala et al., 2006) and t-closeness (Li et al., 2007). Even though these methods provide better privacy preservation for PPDP they result in a higher data utility loss and the above discussed complexities with respect to mixed data, suppressed data and generalized data persist. Therefore, the standard k-anonymity based privacy models are not amenable in the context of ML especially if the underlying data contains categorical QIDs. The main reason for the above discussed complexities of standard k-anonymity based privacy models can be attributed to the requirement of maintaining record level truthfulness for PPDP. In order to satisfy this, data are transformed with truthful methods such as generalization and suppression that reduce the precision of the data but not their accuracy. However, when anonymized data are used for training the supervised ML models (i.e., classifiers)

the record level truthfulness does not play a significant role compared to preserving the relationships between the features and the class attribute.

The above discussion highlights that an anonymization approach that can be easily applicable for mixed data sets, which does not alter the attribute domain or cause data suppression is amenable in the ML context compared to standard k-anonymity or its variants. As a solution, we turn towards the notion of "probabilistic k-anonymity" introduced in Soria-Comas and Domingo-Ferrer (Soria-Comas and Domingo-Ferrer, 2012) which relaxes the indistinguishability requirement of standard k-anonymity while guaranteeing the same level of disclosure risk. A generic framework is introduced in this work for achieving probabilistic k-anonymity concerning numerical data. Unlike the standard k-anonymity, here the indistinguishability requirement is achieved via performing data swapping within the homogeneous data partitions (equivalence classes) of size k. Thus for any given record, its original QID values are dispersed within a group of similar records (with respect to QIDs) of size k leading to uncertainty in record re-identification. Nevertheless, similar to the standard k-anonymity this also limits the probability of re-identification at most to $\frac{1}{k}$. Since this approach does not rely on data suppression and/or generalization it eliminates the aforementioned challenges when adopting into ML context.

In the initial work, it is shown that probabilistic k-anonymity results in better data utility compared to standard k-anonymity. However, these results are limited to the context of numerical data and no proper approach is presented on how data swapping is carried out in order to achieve probabilistic k-anonymity. Most importantly, no analysis has been carried out on how probabilistic k-anonymity would impact disclosure. Nevertheless, data utility evaluation is limited to measuring generic utility losses but no work has been done on its impact when the anonymized data are used for data analysis purposes (i.e. ML). Our contribution in this paper is mainly threefold. First, we present a framework for probabilistic k-anonymity extending it to mixed datasets based on data permutation. Here, we discuss different distance measures appropriate for mixed data in order to generate the homogeneous data partitions for k-anonymity. Secondly, we carry out a comprehensive, empirical evaluation on disclosure risk and data utility with respect to PPDP. Then, we extend the analysis with respect to ML based classification of such data focusing on the impact on the model utility. Finally, we comparatively evaluate our approach with the existing work.

The rest of the paper is organized in the following way. In Section 2 we discuss preliminaries of data anonymization followed by Section 3 detailing the methodology for probabilistic k-anonymity and determination of QIDs. Empirical evaluation and results are presented in Section 4. Section 5 concludes the paper with some final remarks.

2 RELATED WORK

Zhang et al. (Zhang et al., 2007) proposed the notion of (k, e)-anonymity, a permutation based approach to deal with numerical sensitive attributes. Here, the data are partitioned into groups containing at least k different sensitive values within a range of at least e. Then the sensitive values are randomly shuffled within each partition. However, this work is focused on answering the aggregate queries about the sensitive attribute and not about publishing privacy preserving data. Re-identification of individuals is still possible on (k, e)-anonymized data as it does not modify the QID values that can lead to disclosure. An anonymization algorithm is proposed by (Eyupoglu et al., 2018) based on the concept of probabilistic anonymity where the data are anonymized utilizing a chaotic function for data perturbation. Then the resulting anonymized data are used to train a set of classifiers followed by evaluating the classification accuracy. The results show that the proposed algorithm achieves a classification accuracy comparable with the benchmark model. Some other work has used standard k-anonymity based privacy models in order to anonymize the training data (Rodríguez-Hoyos et al., 2018), (Herranz et al., 2010), (Wimmer and Powell, 2014). Despite the data distortion the models induced from anonymized data has reported comparable accuracies with the benchmark model. However, non of these works have highlighted the practical challenges of using standard k-anonymity in the ML context or any analysis to understand how comparable accuracies occur despite the utility loss caused by anonymization. Fung et al. (Fung et al., 2005) presents an algorithm for determining a generalized version of the data that aims at maintaining classification utility. The algorithm generalizes a given dataset by specializing it iteratively starting from the most general state. At each iteration, a general value is assigned into a specific value for categorical attributes, or a given interval is split further for continuous attributes based on information gain. Iterative partitioning of the data is repeated until further specialization leads to violation of the required anonymity level. The results indicate comparable ac-

curacy with the benchmark model even for higher anonymity levels. This method can be applied for mixed datasets given that a semantically meaningful taxonomy tree can be generated for the underlying original dataset. Moreover, at the inference phase transformation of the incoming data (feature vectors) are required to map original values into generalized values. k-Anonymity based anonymization methods first generate homogeneous data partitions (equivalence classes) of size k based on the QID values and then apply data generalization and/or suppression within them. Last et al. presented an anonymization algorithm NSVDist (Non-homogeneous generalization with Sensitive value Distribution) which is based on non-homogeneous (k,l) anonymity where k indicates the required minimal anonymity and l indicates the diversity level for the sensitive attribute. Here, generalization is carried out without clustering the data first. Unlike typical anonymization methods, generalization is applied to the sensitive attribute which converts it into frequency distributions. However, standard ML algorithms cannot be applied to the generalized tables obtained via this approach. Therefore, a data reconstruction step is required before training the classifiers. Privacy Preserving Data Mining (PPDM) algorithms are also proposed in the literature which is aimed at anonymization of the data to cater for the specific data mining goals to maximize the accuracy of the data mining results. In this case, PPDM algorithms are tailored to specific data mining algorithms (i.e., decision trees) under the assumption that the exact use of the anonymized data is known to the data controllers beforehand (Agrawal and Srikant, 2000). Recent work has done an empirical analysis on how applying existing privacy models and anonymization methods on the training data impacts the utility and the privacy of the ML models (Senavirathne and Torra, 2020). Privacy of the ML models is determined based on the success ratio of the membership and attribute inference attacks that target the ML models trained on the anonymized data. Based on their empirical results they have shown that in order to minimize the privacy risks in ML (specifically membership inference), the existing data anonymization techniques have to be applied with high privacy levels that can cause a deterioration in the model utility.

3 DATA ANONYMIZATION

The conventional requirement for data anonymization is to support PPDP. This is a ubiquitous practice adopted in a wide variety of domains where personal data are modified using anonymization techniques and then released publicly or to a specific third party for further analysis (e.g., statistical agencies, research institutes, private/ public organizations). Data anonymization achieves privacy as it irrecoverably transforms the data to minimize the identity and attribute disclosure risks which are respectively aimed at re-identifying individuals and learning previously unknown, confidential characteristics about them.

Based on the impact on privacy we can identify four types of attributes in a given dataset. "Identifiers" are the attributes that can be used to identify the respective data subjects directly e.g., social security number, email address etc. "Quasi identifiers (QIDs)" are indirect identifiers. When QIDs are considered as a composite key they can be used to identify some data subjects accurately. Usually, QIDs are empirically decided by the domain experts who have an extensive understanding of the data. "Confidential/ sensitive attributes" contain sensitive information about data subjects such as health condition, salary, sexual orientation etc. A confidential attribute can also be a QID. "Non confidential attributes" do not contain sensitive information thus no privacy impact is noted. In the process of data anonymization, first, the identifier attributes are removed from the data and then the identified QIDs are modified using appropriate anonymization techniques to produce a protected dataset. In this case, the concepts of "anonymization techniques" (masking methods/ SDC techniques) and "privacy models" play a crucial role. Anonymization techniques direct how to transform the original data into a protected version. In contrast, a privacy model presents a specific condition that, if satisfied guarantees a degree of privacy that keeps disclosure risk under control. Both of these concepts are parametrized and allow the data controllers to tune the degree of privacy that indicates how much disclosure risk is acceptable. There is a synergy between anonymization techniques and privacy models. That is anonymization techniques are used to achieve specific privacy models as they determine how the original data should be transformed. The privacy model k-anonymity (Samarati, 2001) limits the risk of re-identification by ensuring that for a given record there exist at least $k-1$ records that share identical values for QIDs.

Definition 1. *(k-Anonymity) A micro-data set T' is said to satisfy k-anonymity if, for each record $t \in T'$, there are at least $k-1$ other records sharing the same values for all the QIDs.*

Such k records are known as an equivalence class. k-Anonymity decreases the probability of a successful record linkage based on any subset of QID to be

at most $1/k$. k-Anonymity can effectively mitigate identity disclosure as it makes a given data record indistinguishable among $k-1$ other records with respect to QID values. However, attribute disclosure is possible on k-anonymized data if the values for the sensitive attribute are the same or very similar within an equivalence class. In that case, the adversary can infer the sensitive attribute value without prior re-identification. These are known as homogeneity attacks and similarity attacks respectively. In order to avoid the vulnerabilities in k-anonymity some enhancements are proposed. Out of them the most widely known methods are l-diversity (Machanavajjhala et al., 2006) and t-closeness (Li et al., 2007). However, these models are often criticized for their unrealistic assumptions on the sensitive attribute distribution and utility loss. Moreover, l-diversity is also vulnerable to adversarial attacks aimed at learning the sensitive attribute values like similarity attack, background knowledge attack and skewness attack (Li et al., 2007).

With respect to the above discussed k-anonymity based privacy models, the generation of equivalence classes makes data records indistinguishable among k records thus limiting the risk of identity disclosure. However, due to equivalence classes, it becomes easier for an adversary to filter out the exact set of records (i.e., the particular equivalence class) that corresponds to the data record at the adversary's hand. Hence, group identification makes a k-anonymized dataset vulnerable to attribute disclosure (i.e., through homogeneity attack, similarity attack, skewness attack etc.). Whereas, t-closeness is resilient to such attacks at the expense of data utility. This raises the requirement for privacy models focused on improving adversary's uncertainty in correctly identifying the groups (i.e., minimizing attribute disclosure) while maintaining record indistinguishability among a group of records (i.e., minimizing identity disclosure). Therefore, a privacy model becomes more preferable if it introduces uncertainty in both group identification and record re-identification thus minimizing the overall risk of disclosure. These objectives can be achieved by introducing randomness into the privacy model while maintaining a high degree of symmetry within the equivalence classes. By definition, a probabilistic approach for k-anonymity can address these requirements. This notion is referred to as probabilistic k-anonymity and defined as below (Oganian and Domingo-Ferrer, 2017) (Soria-Comas and Domingo-Ferrer, 2012).

Definition 2. *(Probabilistic k-Anonymity) A published data set T' is said to satisfy probabilistic k-anonymity if, for any non-anonymous external data set E, the probability that an adversary with the knowledge of T', E and the anonymization mechanism M correctly links any record in E to its corresponding record (if any) in T' is at most $1/k$.*

Standard k-anonymity limits the probability of re-identification at most to $\frac{1}{k}$ by ensuring for each record in T', there exist at least $k-1$ other records sharing the same values for all the QIDs. On the other hand, the concept of probabilistic k-anonymity relaxes the indistinguishability requirement of standard k-anonymity and only requires that the probability of re-identification be the same as in standard k-anonymity. In the case of probabilistic k-anonymity, indistinguishably is achieved via swapping of the attribute values within the equivalence classes of size k thus creating uncertainty for the adversary in the re-identification process. Even though the definitions of these two privacy models seem to be different from each other at a glance, they enforce the same limit on the probability of re-identification (i.e., $\frac{1}{k}$).

First, from the point of view of privacy, not only probabilistic k-anonymity limits the risk of re-identification/ identity disclosure, it effectively lowers the risk of attribute disclosure as exact group/ equivalence class identification is made difficult via data swapping. With respect to the utility of the anonymized data, probabilistic k-anonymity can preserve the marginal distributions exactly when the entire dataset is concerned in a univariate manner (attribute wise). Also, it permits us to maintain the variability in the anonymized data set as opposed to the standard k-anonymity based methods which reduce the variability via generalization, suppression and/ or aggregation of data that leads to high utility loss (Oganian and Domingo-Ferrer, 2017). Moreover, when the number of selected QIDs are high (e.g., when all the attributes are considered as QIDs) it is shown that standard k-anonymity based methods incur significant utility loss (Aggarwal, 2005). On the other hand, probabilistic k-anonymity distorts the multivariate relationships in the data (i.e., correlations, mutual information etc.). Also, it adversely impacts the analysis done on data sub-domains (e.g., computation of mean salary with respect to a specific job). Therefore, partitioning a given dataset based on their homogeneity before applying data swapping is important as it leads to reduce the above mentioned negative impact on the data utility since the attribute values are now shuffled in a controlled setting. Apart from the above issues, the possibility of unusual combinations of data may occur due to data swapping. To overcome this, data swapping can be carried out in a multivariate manner where first we group several QIDs into a single block and then swapping is applied to each block separately

instead of targeting a single QID at a time.

Apart from the aforementioned privacy and utility related advantages, probabilistic k-anonymity has many favourable characteristics when applied in the ML context compared to the other privacy models. All k-anonymity based syntactic privacy models (i.e., standard k-anonymity, l-diversity, t-closeness etc.) bring up the challenges discussed in Section 1 with respect to data suppression and generalization. Some other data anonymization methods, as discussed in Section 2 either publish anonymized data tailored to specific ML algorithms or require data regeneration in order to transform them into a flexible format before using them for model training. As probabilistic k-anonymity is a general framework focused on controlling the maximum re-identification probability it provides data controllers more flexibility in selecting the underlying data transformation techniques. Therefore, data anonymized via probabilistic k-anonymity can be directly used for model training and inference without any additional requirement for data preprocessing.

4 METHODOLOGY

In this Section, we extend the initial work of Soria-Comas and Domingo-Ferrer (Soria-Comas and Domingo-Ferrer, 2012) with respect to mixed data and present an algorithm for achieving probabilistic k-anonymity based on data permutation. The proposed method consists of four main steps as a) computing pairwise dissimilarity among the data records, b) data partitioning based on QIDs, c) grouping QIDs and, d) data permutation. First, we introduce some basic notions as below. Let $T = \{t_1, t_2, \ldots, t_n\}$ be a dataset with attributes $A_1, \ldots, A_m$. Q denotes the set of all attributes that are considered as QIDs whereas S represents the sensitive attribute. Privacy parameter or the minimum size of each data partition is indicated by k whereas $C = \{c_1, \ldots, c_p\}$ represents data partitions. Dissimilarity measure $d(.)$ is used to measure the pairwise dissimilarity between given two records and finally, a pair-wise dissimilarity matrix (M) is generated.

- **Computing Pairwise Dissimilarity:** Homogeneous data partitions (clusters of size k) are created based on record similarity (or dissimilarity). There exist many similarity measures that can be used to obtain a similarity score. The choice of these measures depends on the data type. For example, in numerical data context Euclidean, Mahalanobi's, Manhattan distances can be used. Whereas, for ordinal and nominal data, measures such as Jaccard's coefficient or Hamming distance are suitable. In their work, Soria-Comas and Domingo-Ferrer have limited their analysis to numerical data and used MDAV-microaggregation (Domingo-Ferrer and Mateo-Sanz, 2002) to create data clusters of size k which is based on the aforementioned numerical distance measures. However, in the real world, most datasets are a mixture of data types hence it is required to use a distance measure appropriate to mixed data. Here, we use two distance measures proposed for mixed data clustering as mentioned below.

Gower dissimilarity - Gower distance (Gower, 1971) is designed to measure the dissimilarity between two data points of mixed data types as depicted below.

$$d_(t_i, t_j) = \frac{\sum_{l=1}^{m} w_{t_i t_j l} d_{t_i t_j l}}{\sum_{l=1}^{m} w_{t_i t_j l}}. \quad (1)$$

Here, $w_{t_i t_j l}$ indicates the weight for variable l between observations i and j, whereas $d_{t_i t_j l}$ indicates the dissimilarity between instance i and j on attribute l. Categorical and numerical attributes are considered separately when calculating $d_{t_i t_j l}$. For categorical data (nominal or binary) distance is measured using Hamming distance where $d_{t_i t_j l}$ is set to 0 when attribute values of i and j are equal and 1 when they are not. For numerical data the scaled absolute difference is used that limits the attribute range between 0 and 1. When computing the weights $w_{t_i t_j l}$ is set to 1 for non-missing variables and 0 otherwise.

Ahmad-Dey dissimilarity - In the above case, Hamming distance is used to evaluate categorical data which is not very accurate for multi-valued categorical attributes. Hence, Ahmad and Dey (Ahmad and Dey, 2011) presented a dissimilarity measure for mixed data that computes a dissimilarity between categorical attribute values based on their co-occurrence with the values of other attributes as mentioned below.

$$d_{t_i, t_j} = \sum_{r=1}^{A_\eta} w_r (t_{ir} - t_{jr})^2 + \sum_{s=1}^{A_\nu} \delta(t_{is}, t_{js}). \quad (2)$$

Here, A_η and A_ν respectively represents numerical and categorical attribute sets. The first term represents the squared Euclidean distance between r^{th} numerical attribute value between instance t_i and t_j. The second term depicts the dissimilarity of the s^{th} categorical attribute value between instances t_i and t_j. $\delta(.,.)$ is used for measuring the dissimilarity between categorical data. Let A_s denote a

categorical attribute which contains two values as a and b. To measure the dissimilarity between a and b, this method considers the overall distribution of a and b in the data set along with their co-occurrence with values of other attributes. Let A_v denote another categorical variable. Let ω denote a subset of values of A_v and $\overline{\omega}$ the complementary set. $P(\omega|a)$ denote the conditional probability that an object having value a for A_s, has a value in ω for A_v. Similarly, $P(\omega|b)$ denotes the conditional probability that an object having value b for A_s, has value in ω for A_v. The dissimilarity between values a and b for A_s concerning A_v is given by $\delta(a,b) = P(\omega|a) + P(\overline{\omega}|b) - 1$, where ω is the subset of values of A_v that maximizes the quantity $P(\omega|a) + P(\omega|b)$. The dissimilarity between a and b is computed with respect to all the other attributes. The average value of dissimilarity is the distance $\omega(a,b)$ between a and b. The significance of the rth numeric attributes are depicted by w_r. To compute w_r, the numerical attributes are first discretized, followed by computing the distances between every pair of discretized values using the same method. Finally the average values are taken as the significance of the attributes. Interested readers are referred to the original publication (Ahmad and Dey, 2011) for more information.

At the end of this step, we generate a dissimilarity matrix M containing the pairwise dissimilarities between all the data instances with respect to QIDs (separately based on equation 1 and 2).

- **Cardinality Constrained Data Partitioning:** In this step data partitioning (clustering) is carried out based on the dissimilarity matrix M such that each data partition contains at least k data instances. To generate cardinality constrained clusters we use MDAV-Microaggregation (Maximum Distance to Average Vector) (Domingo-Ferrer and Mateo-Sanz, 2002). Typically, microaggregation is an SDC method for numerical data protection where the data are first partitioned into micro-clusters of size k followed by replacing them with each micro cluster's centroid value. However, our focus here is only on the data partitioning part thus we update the MDAV-microaggregation algorithm to generate micro-clusters of size k on the dissimilarity matrix M and to return the clusters which contain the corresponding record indices. The output of this step is a nested list (C) that contains micro-clusters each comprises of the record indices that are clustered together based on their similarity as shown by 1.

Algorithm 1: MDAV-microaggregation for clustering records based on their similarity.

Input: M , k
Output: C
while $2k$ *or more rows in* M *remains* **do**
  Randomly select a row q from M
  Find the furthest point p from q
  Select $k-1$ nearest points to p including p and form micro-cluster c_i by fetching their index values
  For all points in c_i remove corresponding rows and columns from M
  if *there are* k *to* $2k-1$ *points left* **then**
    Form a new micro-cluster c_j and fetch their respective index values from M
  else
    Assign the index values of remaining records in M to the last micro-cluster generated
  Append generated micro-cluster/s to form nested list C

- **Generate QIDs Groups:** As explained in Section 3 applying data anonymization in a multivariate manner improves the utility of the anonymized data. This is achieved by grouping QIDs into several blocks before applying anonymization. In this case, we block the QIDs based on their association with the sensitive attribute (S) such that each block contains at least 2 QID attributes. In order to estimate the association, we use Mutual Information (MI). MI measures the association between two random variables capturing both linear and non-linear dependencies. MI between two discrete random variables X and Y can be defined as below.

$$MI(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{3}$$

Here, $H(X)$, $H(Y)$ are the entropy or the uncertainty level of the respective variables that can be measured using $H(X) = -\sum_{x \in X} p(x)\ log\ p(x)$. Whereas, $H(X|Y)$ is the conditional entropy that indicates the amount of uncertainty left in X after observing Y. In Equation 3, the first term explains the entropy of X before Y is known, while the second term indicates the entropy after Y is known. Hence, mutual information is the amount of entropy reduced in X by knowing Y. Therefore, if X and Y are independent of each other the amount of MI is 0 whereas, MI is greater than 0 when X and Y are related. For a given dataset T we quantify the MI between the features and the identified

S as $MI_S = MI(A_v, S)$. MI_S is the mutual information vector that indicates the respective association of each attribute with the S. Then we group the QIDs such that attributes that have high MI are blocked together and each block contains at least 2 attributes.

- **Data Permutation.** Once the data records are clustered based on their similarity the next step is to apply within cluster data permutation in order to achieve probabilistic k-anonymity. As explained previously C contains micro-clusters of record indices. For each $c \in C$ we extract the records from T as $T[c]$ and then randomly permute the order of the QID values followed by updating the permuted values in the original dataset T.

Algorithm 2 summarizes the above mentioned process for generating probabilistically anonymized data.

Algorithm 2: Probabilistic k-Anonymity.

Input: T, Q, S, k
Output: Anonymized dataset : T'
$QID_{df} := T[Q]$
$M :=$ Generate pairwise dissimilarity for QID_{df} `// Algorithm 1`

$C :=$ Generate clusters of size k over M
$P :=$ Generate QID blocks
$T' := T$
for $c \in C$ **do**
  for $p \in P$ **do**
    $EC_{df} := T[c, p]$ `// Extract data chunk given their indices(c) and attributes (in block p )`
    $PM_{df} := EC_{df}$
    while $EC_{df} == PM_{df}$ **do**
      $PM_{df} :=$ Permute (EC_{df})
    $T'[c, p] := PM_{df}$

5 EXPERIMENTAL EVALUATION

In this Section, we use six publicly available datasets for evaluating the anonymized data under the following criteria, a) data utility, b) disclosure risk and, c) impact on ML utility. UCI Contraceptive Methods (1473x9), UCI Mammographic (961x6), UCI Adult (48,842x14), UCI German Credit (863x25), UCI Heart disease (303x14) and, UCI Cardiotocography (1,914x23) datasets are used for experimentation where the class attribute of each dataset is considered as the sensitive attribute(S) (the number of instances and the attributes are mentioned within the parentheses). Selected sensitive attributes are respectively "contraceptive method used", "severity", "income status, "credit rating", "presence of heart disease" and, "presence of cardiac arrhythmia". Respectively UCI Cardiotocography and UCI Contraceptive Methods datasets contain 10 and 3 unique values for the sensitive attributes whereas other datasets have binary sensitive attributes. To avoid the anonymization process being solely impacted by chance, each anonymized dataset is generated in five trials and the results are averaged across them.

To evaluate how differing the number of QIDs impact the data utility and disclosure risk, we have used two approaches for QID selection.

- *Part QID*– Only a subset of attributes are selected as QIDs using ARX anonymization tool (Prasser et al., 2014) which leads to more than 99% of re-identification of the records based uniqueness and separation ratios.
- *Full QID*– All the attributes are considered as QIDs except the sensitive attribute S.

For each of the above mentioned approaches, probabilistic k-anonymity is applied with four privacy/ anonymization levels (k as 5, 50, 100 and 200).

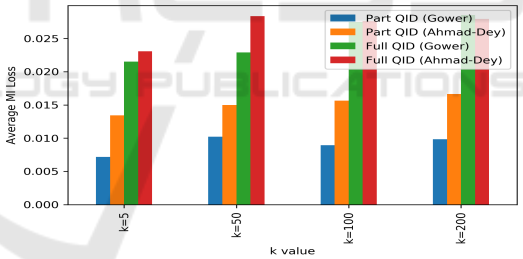


Figure 1: Average loss of mutual information.

5.1 Utility Loss

A generic, data type independent, utility loss measure is used to estimate the utility of the probabilistically k-anonymized data. The measure is based on the concept of mutual information (MI). As explained in (Domingo-Ferrer and Rebollo-Monedero, 2009), mutual information bears some resemblance to the correlation matrix that is used as a generic utility loss measure in SDC literature in the form of the relative discrepancy between correlations. Application of probabilistic k-anonymity alters the statistical dependence between the attributes thus mutual information can be used to quantify this distortion in terms of loss of mutual information. For a given dataset, first, the mutual information matrix is computed which indicates the

association among the attributes. Then the average mutual information vector is computed for each attribute $A_1 \dots A_m$ respectively as $MI_{vec} = MI_1 \dots MI_m$. If the MI_{vec} of the original dataset is termed as MI_T and that of the anonymized dataset is termed as $MI_{T'}$, the relative loss of mutual information is computed as, $MI\ loss = \frac{1}{m}\sum MI_T - MI_{T'}$.

Fig 1 illustrates the probabilistically k-anonymized data under the previously mentioned dissimilarity measures (i.e., Gower, Ahmad-Dey) and QID selection methods (*Part QID*, *Full QID*). The results are averaged across the selected six UCI datasets. Here, we can observe that increasing the anonymization level (k) and/or number of QIDs increase the loss of MI. This can be attributed to the higher distortion caused by data permutation within larger data partitions which leads to higher privacy. Moreover, comparing to tuning the privacy parameter (k) into a higher value, selecting a higher number of QIDs has a more adverse impact on data utility. Therefore, data controllers have to be more economical when selecting QIDs for anonymization. When data partitioning is done based on Gower dissimilarity the relative loss of MI is marginally lower than that of Ahmad-Dey dissimilarity.

5.2 Disclosure Risk

The purpose of disclosure risk limitation is to minimize the amount of information available in an anonymized dataset with respect to individuals leading to identity and/ or attribute disclosure. In the case of probabilistic k-anonymity, an adversary would not be able to positively identify which anonymized records correspond to which original records exactly. In other words, it is impossible for an adversary to be confident of having identified an individual in the anonymized dataset as every record now could be altered. Therefore, the risk of identity disclosure is minimal under probabilistic k-anonymity except in the presence of unique values in QIDs. For example, if a particular record carries unique values for all or most of the QIDs, even if those values are dispersed among multiple records (via permutation) an adversary can still draw a conclusion of its mere existence in the anonymized dataset (without an exact record linkage).

One of the biggest limitations of standard k-anonymity is the vulnerability towards attribute disclosure. In standard k-anonymity, risk of attribute disclosure occurs due to the possibility of exact equivalence class identification followed by the existence of same (or similar) sensitive attribute values within the given equivalence classes (refer Section 3 for more information). In this case, an adversary can obtain the sensitive attribute values of a given record by observation without using any sophisticated inference methods. We use Adult and Cardiotocography datasets to showcase this risk. On the Adult dataset, the sensitive attribute has two unique values whereas on the Cardiotocography dataset ten unique values are available. When standard k-anonymity is applied on the Adult dataset, respectively 0.72%, 0.6%, 0.52%, 0.47%, 0.11%, 0.06%, and 0.002% records belong to equivalence classes where the same sensitive attribute value is present when k value differs as 2, 3, 4, 5, 50, 100 and 200. On the Cardiotocography dataset for the aforementioned k values percentage of records belong to the equivalence classes with the same sensitive attribute value changes as 0.45%, 0.27%, 0.24%, 0.027%, 0%, 0%, 0%. This shows that small privacy parameter values (k) and low diversity in sensitive attributes could result in high attribute disclosure risk under standard k-anonymity.

Even though exact equivalence class identification is not possible with high certainty in probabilistically k-anonymized data attribute disclosure is still possible. However, by definition attribute disclosure is more challenging under probabilistic k-anonymity compared to standard k-anonymity. Hence, to assess this risk we adopt three inference methods namely, a) distance based record linkage b) probabilistic record linkage and, c) ML based inference.

In this case, we consider the scenario where an adversary has access to an external dataset E that comprises un-anonymized QID data and E's anonymized version T'. Here, we assume that E contains personally identifiable information (unmodified QIDs that can lead to specific individuals) without the sensitive attribute/s. Whereas, T' contains the sensitive attribute values along with the modified (anonymized) QIDs. Thus, by using a mechanism to join E and T' on their QIDs adversary can infer the sensitive attribute value for the interested individuals in E. To illustrate a sophisticated adversary with full access to information, we assume E contains all the records included in T', prior to anonymization without the sensitive attribute values (i.e., $E = T \setminus SA$).

Distance based Record Linkage. In this case, the adversary uses record similarity to identify a potential link for a given record in E and infer the sensitive attribute based on this. For a given record $t_i \in E$ the adversary computes the similarity to all the records $t_{i...n} \in T'$. Then for $t_i \in E$ find the nearest neighbour from T' based on the computed similarities and extract the sensitive attribute value. In this case, we have used Gower dissimilarity for generating the distance matrix as it incurs a lower utility loss compared to Ahmad-Dey.

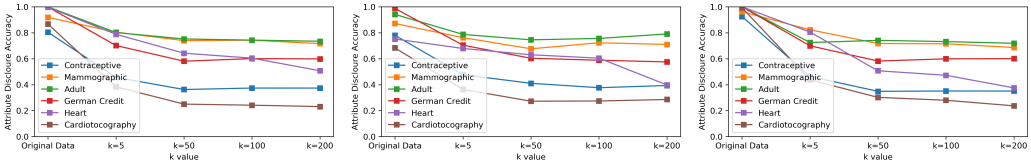
Probabilistic Record Linkage. Probabilistic record linkage (PRL) attempts to link two datasets when there are inconsistencies between the records with respect to the linking attributes. The goal of this record linkage method is to establish whether a given pair of records belong to the same entity or not. In this case, probabilistic record linkage is carried out between E and T' on QID attributes. As probabilistic k-anonymity permutes the QID values in T' now there are inconsistencies between the majority of the linking attributes. Thus even a record is inked, the chances are high that it is merely a similar record but the correct match. To illustrate this we generate a dataset T_v. 50% of T_v's records are from T' and the rest is from a hold out dataset T_h where $T_h \cap T = \emptyset$. Then we used probabilistic record linkage to identify links (based on reclin package on R) where one-to-one linkage is enforced. Ideally, 50% of the records taken from T' should have been correctly liked while the records taken from T_h are not linked. In other words, a high True Positive (TP) count and a low False Positive (FP) count is expected. However, we noticed that probabilistic record linkage results in a very high number of FPs with respect to all the datasets. Thereby, reporting low positive predictive value (PPV) computed as $PPV = \frac{TP}{TP+FP}$. The average of the reported PPV values is 0.49 ± 0.01 across all the UCI datasets for their probabilistically k-anonymized counterparts. This result indicates that probabilistic record linkage under probabilistic k-anonymity does not yield meaningful results. Nevertheless, we can utilize the linked records to infer the sensitive attribute values for records in E.

Once the records are linked between E and T' the next step is to infer the sensitive attribute (S) value as mentioned below. Assume record $t_i \in E$ is linked with $t_j \in T'$. Then value of S is inferred as $E_{t_i}[S] = T'_{t_j}[S]$.

ML based Inference. In this case, the adversary first trains a ML classifier on T' in order to predict the sensitive attribute of the original, un-anonymized dataset E. Since it is assumed that E only contains QIDs, the ML classifiers are trained only based on the QIDs attributes of T' In this work, we have used the random forest (RF) algorithm as the adversary's choice to train the ML model. The accuracy of the inferred SAs is reported using the micro-averaged F1 score which aggregates the contributions of all the classes.

The average AD risk under the aforementioned inference methods are illustrated in Fig 2 when the underlying micro-data are partitioned based on Gower dissimilarity and only a subset of attributes are considered as QIDs (*Part QID*) following the typical adversarial assumption in SDC where it is assumed that the adversaries have access to only a subset of attributes that can be effectively used for re-identification (i.e., QIDs). We term such adversaries as *partially informed adversaries (PIA)*. More precisely, in this case, the dataset E contains only a subset of attributes (un-anonymized). The fraction of attributes (QIDs) available in E concerning each dataset is 0.23, 0.2, 0.44, 0.18, 0.42, 0.8 and, 0.35 respectively for Heart, German Credit, Contraceptive Methods, Cardiotocography, Adult and, Mammographic datasets. On each dataset, the attribute disclosure risk (AD) is first measured on the original, un-anonymized dataset followed by an evaluation of the probabilistically k-anonymized data. The purpose of using the original, un-anonymized dataset in the experimentation is purely to evaluate the effectiveness of the different attribute inference methods. With respect to all the inference methods, it can be seen that AD accuracy is very high for original, un-anonymized data. That is on average 0.92, 0.83 and, 0.98 for distance based, ML based and, PRL based methods respectively. Considering all the datasets it can be seen that applying probabilistic k-anonymity has reduced the risk of AD. For varying privacy levels (i.e., k=5, 50, 100, 200) on average, AD risk is reduced by ≈ 39.9 %, ≈ 25 % and ≈ 49 % respectively for for distance based, ML based and, PRL based methods. However, AD risk is not always monotonically reduced as we increase the privacy level k. For example, Cardiotocography and Contraceptive datasets seem to be more susceptible to increasing k values compared to the other datasets which have binary sensitive attributes (class attribute). However, datasets with binary class attributes have also shown a reduction of AD approximately about 20% indicating that probabilistic k-anonymity result in reducing AD risk effectively.

Next, we summarize the results of AD risk when all the attributes are treated as QIDs ($Full\ QID$). In this case, we assume that the external dataset E, that the adversary has access to, contains all the attributes except for the sensitive attribute S. We term such adversaries as *fully informed adversaries (FIA)*. FIA can be considered as a real threat since an adversary would at least have the complete knowledge of the data belong to his/her close contacts (i.e., family, friends). Or with the excessive digital data collection, the availability of such personal data is no longer a far fetched assumption. Here, the dataset E contains all the attributes (un-anonymized) and the adversary exploits them to infer the correct sensitive attribute value based on the aforementioned inference methods. The AD risk reduction is ≈ 38 % for distance based approach, ≈ 22 % for ML based approach, ≈ 51 % for PRL based approach. These re-



(a) Distance Based (b) ML Based (c) PRL Based

Figure 2: Attribute disclosure risk under Gower distance and *Part QID*.

sults show that probabilistic k-anonymity can reduce the AD risk successfully in the presence of strong adversaries with a significant amount of background knowledge about the underlying data subjects. (Similar results are also noted for Ahmad-Dey dissimilarity. They are not included here due to space limitations.) The take away from these results is two-fold. First, probabilistic k-anonymity can effectively reduce AD risk. As explained earlier, AD is possible in standard k-anonymity due to exact equivalence class identification which is no longer exists in probabilistic k-anonymity. Even with sophisticated inference methods, fully informed adversaries AD risk remains low when data are probabilistically k-anonymized. Secondly, probabilistic k-anonymity cannot completely alleviate AD risk. As long as the data remains useful there exists room for AD.

5.3 Impact on Machine Learning

In the beginning, we discussed why probabilistic k-anonymity is an amenable approach in the context of ML compared to the rest of the well-known privacy models used for data anonymization. In this sub-section, we discuss the impact of training data anonymization based on probabilistic k-anonymity with respect to ML model utility.

Model Utility. Fig 3 depicts the F1 score when the underlying training data are anonymized using probabilistic k-anonymity. Here, we have used two QID selection methods as *Part QID* and *Full QID* which varies on the number of QIDs. As shown by the results, increasing the number of anonymized QIDs and the privacy levels (k) deteriorate the model accuracy gradually. Previously, we discussed how probabilistic k-anonymity incurs utility loss in the data and used the loss of MI to quantify it. Hence, it is intuitive that the ML models trained on the anonymized data also face a loss of predictive power. Here, the F1 score is measured on a holdout dataset (T_h) extracted from the same population as T without any overlapping. Multi Layer Perceptron (DNN), Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT) and Random Forest (RF) algorithms are used in the experiments and the reported accuracy values are averaged over them. For DNN a three layer neural network is configured with 32 hidden units each with "relu" activation and "adam" optimizer.

However, we do not observe linearity in the accuracy loss as a response to the increasing number of QIDs or privacy level. Further, in some cases, the loss of model accuracy is almost negligible or even slightly improved despite the utility loss caused by anonymization. Concerning the different QID selection methods, the average F1 score loss varies from 0.12 ± 0.01 to 0.23 ± 0.05 under Gower dissimilarity concerning *Part QID* and *Full QID* respectively. With respect to Ahmad-Dey dissimilarity the average F1 score loss changes as 0.15 ± 0.01 and 0.24 ± 0.01. In a closer inspection of the results, it can be seen that binary classification problems are less susceptible to utility loss compared to the multi-class problems in general. This is more prominent under *Full QID* approach for QID selection. As explained by Senavirathne and Torra (Senavirathne and Torra, 2020) in multi-class classification, only a limited number of records per class exist for the classification algorithm to learn a discriminative pattern. When anonymization distorts the existing relationships in the data it becomes increasingly difficult to learn an accurate pattern. This explains the high accuracy loss in multi-class cases. In order to improve the model accuracy in multi-class cases, we can balance the class distribution and/ or re-define the classification problems to have a limited number of classes when it is possible. From the above results, it is conspicuous that the use of probabilistically k-anonymized training data impacts the classification accuracy negatively. By opting lower privacy level (k) and/ or a smaller subset of QIDs this negative impact can be lowered. However, data controllers have to keep in mind that tuning for more accuracy leads to high disclosure risk.

Comparative Analysis. In this Section, we compare probabilistic k-anonymity with other commonly used privacy models and comparatively evaluate their impact on classification utility with respect to Deep Neural Networks (DNN) as they are being increasingly used to solve complex ML problems. For the evaluation, we use standard syntactic privacy mod-

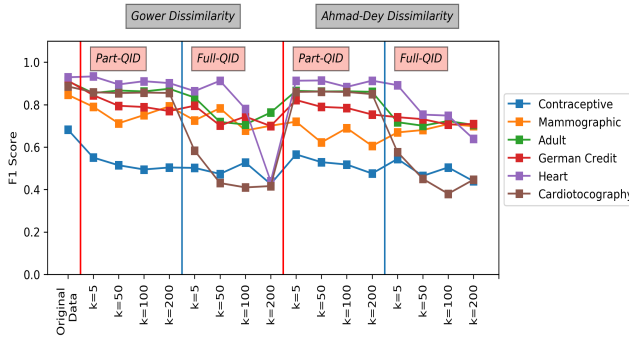


Figure 3: Average F1 Score.

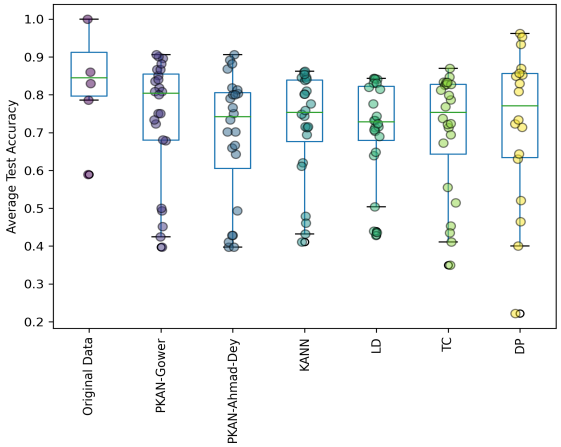


Figure 4: Comparative evaluation of test accuracies for DNNs trained under different privacy models. Each marker indicates a different dataset and a different privacy level.

els such as k-anonymity (KN), l-diversity (LD) and t-closeness (TC). In this case, the anonymized data are generated based on Mondrian anonymization algorithm (LeFevre et al., 2006). As we discussed earlier, learning based on differentially private (DP) ML algorithms have become the standard practice for PPML. Hence, we adopt DP DNN (Abadi et al., 2016) for the comparative analysis as well.

As explained at the beginning, compared to probabilistic k-anonymity (PKAN) the standard syntactic privacy models require the support of data transformation APIs in the inference phase of ML as they alter the attribute domain of the underlying data. DP also faces a myriad of challenges including high utility loss when applied to ML. However, amongst those challenges having to maintain access to raw, sensitive data in order to train the DP ML models greatly hinders the flexibility of sensitive data analysis (Refer Section 1 for more details). Compared to these methods probabilistic k-anonymity not only provides straight forward implementation also provides high flexibility for sensitive data processing without risking compliance with GDPR.

Figure 4 showcases the test accuracy obtained via each privacy model over aforementioned datasets when only a subset of attributes are considered as QIDs (*Part QID*). A one to one comparison between these privacy models is not very meaningful as there are differences between how each privacy model is implemented and what the privacy parameters mean in each case. However, our attempt here is to understand if each of these privacy models is implemented with acceptable privacy levels to generate anonymized data, how would it impact the classification accuracy of the ML models induced from that? To realize this objective we train multiple ML classifiers for each dataset with a variety of privacy levels. For probabilistic k-anonymity and standard k-anonymity privacy levels (k) are chosen as 5, 50, 100, and 200. For t-closeness privacy levels (t) are chosen as 0.5, 0.3, 0.1, and 0.01. When generating l-diversity data for all the datasets with binary class attributes l value is set to 2. For Contraceptive Methods dataset, l value is set to 2 and 3 respectively. For the Cardiotocography dataset privacy parameter l is set to 2, 3, 4, and 5 respectively. For training differentially private ML classifiers privacy parameter ε is chosen as 3, 1, 0.5, and 0.25 with $\delta = 1e-5$. Once the ML classifiers are trained for each dataset and privacy level, their test accuracies are obtained and grouped over different privacy models to comparatively analyse their impact on ML utility. As depicted in Figure 4 there is a loss of utility caused by applying anonymization which is shown through low accuracy compared to the benchmark model (model trained on original data). Probabilistic k-anonymity implemented with Gower distance shows a higher accuracy compared to probabilistic k-anonymity implemented with Ahmad-Dey distance, standard syntactic privacy models (KAN, LD, TC) and DP. In conclusion, probabilistic k-anonymity obtain a relatively high utility for ML while providing the data controllers with the previously discussed advantages such as high flexibility for sensitive data analysis under GDPR, a means for PPDP with low attribute disclosure risk and, an easy adaptation into ML context without additional data pre-processing or post-processing requirements.

6 CONCLUSION

In this work, we systematically show that probabilistic k-anonymity can effectively address the challenges faced by standard privacy models in the context of ML. Here, we have presented a framework that consists of two algorithms for obtaining proba-

bilistically k-anonymized data in the context of mixed datasets. Then an in-depth analysis is carried out to evaluate the utility and privacy aspects of probabilistic k-anonymity with respect to PPDP. Then we trained a variety of ML classifiers on probabilistically k-anonymized data and evaluated the model utility. When applied with high privacy parameter levels(k) or a high number of QIDs, probabilistic k-anonymity has an adverse impact on ML utility. However, compared to the other syntactic privacy models (i.e., k-anonymity, l-diversity, t-closeness) probabilistic k-anonymity has gained better ML utility. In conclusion, probabilistic k-anonymity obtain relatively high utility for ML while providing the data controllers with numerous advantage such as high flexibility for sensitive data analysis under GDPR, a means for PPDP with low attribute disclosure risk and, an easy adaptation into ML context without additional data pre-processing or post-processing requirements. In future work, it can be explored whether these classification accuracies can be improved further via noise correction and sample selection methods presented in the ML literature when learning has to be carried out on the noisy data.

ACKNOWLEDGMENT

This work is supported by Vetenskapsrådet project: "Disclosure risk and transparency in big data privacy"(VR 2016-03346 , 2017-2020).

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM.

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, pages 901–909. VLDB Endowment.

Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 439–450.

Ahmad, A. and Dey, L. (2011). A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. *Pattern Recognition Letters*, 32(7):1062–1069.

Al-Rubaie, M. and Chang, J. M. (2019). Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy*, 17(2):49–58.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and data Engineering*, 14(1):189–201.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.*, 14(1):189–201.

Domingo-Ferrer, J. and Rebollo-Monedero, D. (2009). Measuring risk and utility of anonymized data using information theory. In *Proceedings of the 2009 EDBT/ICDT Workshops*, pages 126–130.

Eyupoglu, C., Aydin, M. A., Zaim, A. H., and Sertbas, A. (2018). An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, 20(5):373.

Fung, B. C., Wang, K., and Yu, P. S. (2005). Top-down specialization for information and privacy preservation. In *21st international conference on data engineering (ICDE'05)*, pages 205–216. IEEE.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

Herranz, J., Matwin, S., Nin, J., and Torra, V. (2010). Classifying data from protected statistical datasets. *Computers & Security*, 29(8):875–890.

Ji, Z., Lipton, Z. C., and Elkan, C. (2014). Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering (ICDE'06)*, pages 25–25. IEEE.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24. IEEE.

Oganian, A. and Domingo-Ferrer, J. (2017). Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion. *Transactions on data privacy*, 10(1):61.

Prasser, F., Kohlmayer, F., Lautenschläger, R., and Kuhn, K. A. (2014). Arx-a comprehensive tool for anonymizing biomedical data. In *AMIA Annual Symposium Proceedings*, volume 2014, page 984. American Medical Informatics Association.

Rodríguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Parra-Arnau, J., and Forné, J. (2018). Does k-anonymous microaggregation affect machine-learned macrotrends? *IEEE Access*, 6:28258–28277.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

Senavirathne, N. and Torra, V. (2020). On the role of data anonymization in machine learning privacy. In *2020 IEEE 19th International Conference on Trust, Secu-*

rity and Privacy in Computing and Communications (TrustCom), pages 664–675. IEEE.

Soria-Comas, J. and Domingo-Ferrer, J. (2012). Probabilistic k-anonymity through microaggregation and data swapping. In *2012 IEEE International Conference on Fuzzy Systems*, pages 1–8. IEEE.

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588.

Wimmer, H. and Powell, L. (2014). A comparison of the effects of k-anonymity on machine learning algorithms. In *Proceedings of the Conference for Information Systems Applied Research ISSN*, volume 2167, page 1508.

Zhang, Q., Koudas, N., Srivastava, D., and Yu, T. (2007). Aggregate query answering on anonymized tables. In *2007 IEEE 23rd international conference on data engineering*, pages 116–125. IEEE.