Improving Accuracy of Classification Models Induced from Anonymized Datasets

Mark Last[a], Tamir Tassa[b], Alexandra Zhmudyak[a], Erez Shmueli[a]

[a] *Department of Information Systems Engineering, Ben-Gurion University of the Negev, Israel*
[b] *Department of Mathematics and Computer Science, The Open University, Israel*

Abstract

The performance of classifiers and other data mining models can be significantly enhanced using the large repositories of digital data collected nowadays by public and private organizations. However, the original records stored in those repositories cannot be released to the data miners as they frequently contain sensitive information. The emerging field of Privacy Preserving Data Publishing (PPDP) deals with this important challenge. In this paper, we present NSVDist (Non-homogeneous generalization with Sensitive Value Distributions) — a new anonymization algorithm that, given minimal anonymity and diversity parameters along with an information loss measure, issues corresponding non-homogeneous anonymizations where the sensitive attribute is published as frequency distributions over the sensitive domain rather than in the usual form of exact sensitive values. In our experiments with eight datasets and four different classification algorithms, we show that classifiers induced from data generalized by NSVDist tend to be more accurate than classifiers induced using state-of-the-art anonymization algorithms.

Keywords: Privacy Preserving Data Publishing, Privacy Preserving Data Mining, k-Anonymity, ℓ-Diversity, Non-homogeneous Anonymization, Classification

1. Introduction

A vast amount of information of all types is collected daily about people by governments, corporations and individuals. As a result, there is an enormous quantity of privately-owned records that describe individuals' finances, interests, activities, and demographics. These records often include sensitive data and may violate the privacy of the users if published. This information is becoming a very important resource for many systems and corporations that may enhance and improve their services and performance by inducing novel and potentially useful data mining models. One common practice for releasing such confidential data without violating privacy is applying regulations, policies and guiding principles for the use of the data. Such regulations usually entail data distortion operations such as generalization or random perturbations. The challenge with this approach is that, on one hand, data leakage can still occur, and, on the other hand, the data and the resulting data mining models may become nearly useless after excessive distortion [7].

The emerging research field of Privacy Preserving Data Publishing (PPDP) is targeting this challenge [7]. It aims at developing techniques that enable publishing data while minimizing distortion for maintaining utility on one hand, and ensuring that privacy is preserved on the other hand. In this paper we present a new privacy-preserving data publishing method, which is shown to maintain the predictive utility of supervised classification algorithms that are trained on the published data. The predictive utility is measured by the classification accuracy of the induced classification models, when applied to new, previously unseen data. As we explain in the related work section (Section 2), we assume that the validation data can be kept in its original non-distorted form.

A closely related research area is Privacy Preserving Data Mining (PPDM) that was initiated in 2000 by [1]. PPDM algorithms aim at anonymizing data towards its release for specific data mining goals, so that the data utility is maximized, on one hand, and its privacy is preserved on the other hand. The developed PPDM algorithms are tailored to specific data mining tasks and algorithms. For example, if the data needs to be used for inducing a decision-tree classifier, the corresponding PPDM algorithm will aim at achieving anonymization while incurring a minimal loss of accuracy in the resulting classifier. In PPDP, on the other hand, the exact purposes of the data release are unknown and it is needed to anonymize the data using utility measures that are not targeted to a specific data mining algorithm.

It is customary to distinguish between four types of attributes in the database table that needs to be published (see [3]):

- Identifiers — attributes that uniquely identify an individual (e.g. `name`);
- Quasi-identifiers — publicly-accessible attributes that do not identify a person, but some combinations of their values might yield unique identification (e.g., `gender`, `age`, and `zipcode`);
- Sensitive information — attributes of private nature, such as medical or financial data (in this paper, we follow the common assumption of a single sensitive attribute, which is identical to the class attribute); and
- Other non-sensitive attributes that, on one hand, cannot be used for identification since they are unlikely to be accessible to the adversary, and, on the other hand, do not represent information of sensitive nature. (Those attributes can be ignored in our discussion.)

A common practice in PPDP and PPDM is to remove the identifiers and to generalize or suppress the quasi-identifiers in order to protect the sensitive data of individuals from being revealed. Generalization means that the original values of quasi-identifiers are replaced with less specific values, whereas in case of suppression no values are released at all. The sensitive data is usually retained unchanged.

In the past years, several models were suggested for maintaining privacy when disseminating data. Most approaches evolved from the basic model of k-anonymity [38]. In that model, the practice is to remove the identifiers and generalize the quasi-identifiers as described above, until each generalized record is indistinguishable from at least $k - 1$ other generalized records, when projected on the quasi-identifiers. Consequently, an adversary who wishes to trace a record of a specific person in the anonymized table, will not be able to trace that person's record to subsets of less than k anonymized records.

As an example, consider the basic table in Table 1(a), having the quasi-identifiers `Age` and `Zipcode` and the sensitive attribute `Disease`. Table 1(b) is a corresponding 2-anonymization. (Here "M" is short for "Measles", "F" is short for "Flu" and so forth.) An adversary who wishes to trace Eve's record in it may infer that it is one of the last two records, but they are equally likely, whence the probability of correct identification is 1/2. Many algorithms were suggested in the literature for k-anonymization, e.g. [2, 10, 11, 13, 19, 24, 25, 35, 36, 39].

`Name`	`Age`	`Zipcode`	`Disease`
Alice	30	10055	Measles
Bob	21	10055	Flu
Carol	21	10023	Angina
David	55	10165	Flu
Eve	47	10224	Diabetes

(a) The original table

`Age`	`Zipcode`	`Dis.`
21-30	100**	M
21-30	100**	F
21-30	100**	A
47-55	10***	F
47-55	10***	D

(b) Homogeneous anonymization

`Age`	`Zipcode`	`Dis.`
21-30	10055	M
21	100**	F
21-30	100**	A
47-55	10***	F
47-55	10***	D

(c) Non-homogeneous anonymization

Table 1: A table and corresponding anonymizations

The k-anonymity model on its own does not provide a sufficient level of privacy. Its main weakness is that it does not guarantee sufficient diversity in the sensitive attribute within each equivalence class (or block) of records that are indistinguishable by their generalized quasi-identifiers. Namely, even though it guarantees that every record in the anonymized table is indistinguishable from at least $k-1$ others, it is possible that the distribution of the sensitive values in those records discloses "too much" information. To mitigate this problem, Machanavajjhala et al. [28] proposed the security measure of ℓ-diversity. That measure requires that each block of indistinguishable records will have at least ℓ "well represented" sensitive values. One of the interpretations of ℓ-diversity [42, 44] requires that the relative frequency of each of the sensitive values within each block is at most $1/\ell$. The 2-anonymization in Table 1(b) satisfies also 2-diversity. Other measures limiting the information leaked by the distribution of the sensitive attribute in each block are t-closeness [27] and p-sensitivity [40]. A common thread in all those privacy models is that the table records are first clustered into clusters that are required to satisfy some privacy condition, and then all records in a given cluster are replaced with the least generalized record that generalizes all of them.

Gionis et al. [11, 39] proposed a novel approach that suggests achieving anonymity without clustering. In their approach, k-anonymity is achieved by generalizing the table records until each original record can be linked with at least k generalized records, but there is no requirement that each generalized record will have at least $k - 1$ other generalized records that agree with it in their quasi-identifiers. Similarly, ℓ-diversity is achieved by generalizing the table records to the extent that no original record can be linked to any of the sensitive values with probability greater than $1/\ell$. They showed that by breaking out of the clustering paradigm, it is possible to achieve similar levels of anonymity with smaller information losses. The recent study [43] further explored that idea and suggested the term *non-homogeneous anonymization* for such non-cluster based anonymizations. Table 1(c) is a non-homogeneous 2-anonymization of the original table. It may be verified that even an adversary who knows the quasi-identifiers of all records in Table 1(a) cannot link any such record with any of the generalized records in Table 1(c) with probability greater than 1/2. In addition, it can be shown that such an adversary cannot use Table 1(c) to infer links between any of the records in Table 1(a) with any disease with probability greater than 1/2. As Table 1(c) involves less data distortion than Table 1(b), non-homogeneous anonymization can achieve similar privacy goals as homogeneous anonymization with less information loss.

The studies [11, 39, 43] proposed algorithms for achieving non-homogeneous anonymizations and demonstrated the advantage that they offer, compared to homogeneous anonymization algorithms, in terms of information loss. All studies thus far that considered homogeneous or non-homogeneous anonymizations assumed that only the quasi-identifiers are subjected to generalization, while the sensitive attribute remains unchanged. In this paper, we extend the non-homogeneous anonymization framework by allowing the generalization of the sensitive column as well. The generalization is performed in a new way by replacing the sensitive values with frequency distributions over the sensitive domain. We show empirically that such anonymizations enable to learn more accurate classifiers from anonymized data.

Originality and contribution. In the first part of this work we describe NSVDist, a new anonymization algorithm that, given minimal anonymity and diversity parameters (k and ℓ) along with an information loss measure, issues corresponding non-homogeneous anonymizations where the sensitive column is published as frequency distributions over the sensitive domain rather than exact sensitive values, as in the case of customary generalizations. In the second part, we demonstrate the advantages offered by such anonymizations. Previous studies [11, 39, 43] have shown that non-homogeneous anonymizations result in lower information losses than homogeneous anonymizations for the same values of k and ℓ. Those findings raise the question whether such non-homogeneous anonymizations of training data tables improve the utility of induced data mining models on new (validation) data. Focusing on the task of classification, we first explain how to prepare generalized tables so that they can be processed by standard classification algorithms. Then, we show empirically that classifiers that are built using NSVDist tend to be more accurate than classifiers that are built by state-of-the-art anonymization algorithms. In addition, we show that the maximum values of the security parameter k that allow induction of meaningful classification models from the anonymized data are considerably higher with our algorithm (NSVDist) than with state-of-the-art algorithms of standard anonymization.

Organization of the paper. In Section 2 we review related work on privacy-preserving data publishing. In Section 3 we present our extended generalization framework based on the non-homogeneous generalization paradigm. Our algorithm for Non-homogeneous generalization with Sensitive Value Distribution (NSVDist) is introduced in Section 4. The proposed generalization methodology is evaluated in Section 5. Section 6 concludes with a discussion of results and proposed directions for future research.

2. Related work

Fung et al. [8] present a privacy-preserving data publishing method that aims at maintaining classification utility. The proposed Top-Down Specialization (TDS) algorithm performs an iterative top-down partition of the data taxonomy tree as long as the anonymity requirement is preserved and at least two distinct sensitive values are involved in the records containing the specialized domain value. The best specialization is found at each iteration using the well-known information gain measure. The method is evaluated on the `Adult` dataset with C4.5 and Naïve Bayes classifiers.

LeFevre et al. [26] provide a suite of anonymization algorithms that produce a new anonymous view of the given table for each pre-defined set of workloads, consisting of one or more specific data mining tasks, as well as selection predicates. Their approach does not agree with the "non-expert data publisher" assumption [7] according to which many data owners do not have expertise in data mining and they are interested to publish their data only once (e.g., on the UCI Repository) for an unrestricted use by the data mining community rather than for specific data mining tasks.

In [9], the authors propose a k-anonymization solution for classification. The goal is to find a k-anonymization, not necessarily optimal in the sense of minimizing information loss, that retains useful information for classification. That study assumes that the data miner is interested in estimating the testing accuracy on anonymized data, which does not necessarily represent a typical privacy-preserving data publishing situation.

A privacy model called LKC-privacy for anonymizing high-dimensional data along with a top-down specialization Privacy-Aware Information Sharing (PAIS) algorithm are presented in [32]. LKC-privacy upper-bounds the probability of a successful identity linkage by $1/K$ and the probability of a successful attribute linkage by C, provided that the adversary's prior knowledge is limited to at most L of the quasi-identifier values. (For example, (α, k)-anonymity [42] is a special case of LKC-privacy where L is the overall number of quasi-identifiers, $K = k$, and $C = \alpha$.) The PAIS algorithm applies homogeneous anonymization, and it uses two utility measures: The first one (InfoGain) preserves the maximal information for classification analysis. The second one (the discernibility cost measure) aims at minimizing the overall data distortion; it is intended for use when the data mining task is unknown to the data anonymizer. Their algorithm is evaluated on the `Adult` and `Blood` datasets with the C4.5 classifier.

Mohammed et al. [31] propose a generalization-based anonymization algorithm for the so-called non-interactive setting. In that setting, which is assumed by most studies, including ours, the database owner first anonymizes the raw data and then releases the anonymized version for public use. This setting is different from the interactive one, where the data miner is allowed to pose aggregate queries to the database. The solution proposed in [31] first probabilistically generalizes the raw data and then adds noise to guarantee ε-differential privacy. They showed that data generalized in that manner can be used effectively to build a specific decision-tree induction algorithm (C4.5).

Kisilevich et al. [21] propose a new method for achieving k-anonymity without the need for manually producing domain hierarchy trees. Their method, called k-Anonymity of Classification Trees Using Suppression (kACTUS), identifies attributes that have less influence on the classification of the data records; those attributes are then suppressed until the table becomes k-anonymized. Their approach assumes that the data owner is capable of performing data mining on her/his private data, in order to identify the attributes with smaller impact on classification; thus, it is inconsistent with the prevailing assumption of the "non-expert" data publisher who does not have the knowledge needed for running data mining algorithms.

Iyengar [18] uses a genetic algorithm to find an optimal homogeneous generalization of a given dataset in terms of two information loss measures: a general loss metric (LM) and a classification metric (CM). In his evaluation, he also assumes that the data miner is interested in applying the induced model on the anonymized data, which may be generalized and published in several releases. His results on the `Adult` dataset indicated that with the CM metric there was little increase (up to 1.4%) in the error rate as the privacy requirement ranged from $k = 10$ to $k = 250$. It is noteworthy that our algorithm (NSVDist) exhibited the same level of accuracy loss only for $k = 400$.

Rather than using user-defined domain generalization hierarchies, Nergiz and Clifton [33] present a family of clustering-based generalization algorithms. They argue that anonymization quality metrics strongly depend on the intended data mining task, and if that task is known in advance, the data owner can simply release the appropriate model (e.g., a classifier) instead of risking a privacy breach by publishing anonymized data. According to their experimental results, no single information loss metric can be used as a reliable predictor of data mining performance. (In this study we found that NSVDist, which is characterized by smaller information losses than other anonymization algorithms, does perform better in terms of classification accuracy.)

Herranz et al. [17] evaluate the utility of several Statistical Disclosure Control (SDC) methods for constructing accurate classifiers from protected data. The induced models are used to classify future records that are assumed to be given in their original form, without any protection. The same assumption is implemented in our work. All anonymization methods used by [17] are limited to numeric attributes and they do not take into account the

ℓ-diversity constraint. The results of the experiments in [17] indicate that the performance remains essentially unchanged for the lower levels of protection, and it degrades slowly as the level of protection grows.

In view of the real-world data publishing constraints, we did not choose to follow the paradigm of [8] and [32]. Their approach is to split the published anonymized data into two parts – the training set (used for model induction) and the testing set (used for model evaluation). As indicated above, this evaluation approach does not represent a typical "non-expert data publisher" scenario [7], where the data miners are primarily interested in applying the model induced from the published data to their own private data, which does not have to be published or anonymized. Thus, in our research, we have anonymized only the training set, while the records from the testing set were used for classification in their original (non-anonymized) form.

We conclude this review of related work by noting that the idea that some attributes may not be essential for classifying specific database objects goes back to Kryszkiewicz [22, 23]. The *reduct* of a database table is defined by Kryszkiewicz as a minimal subset of attributes required for identifying a given object with certainty. Such subsets of attributes relate to the notion of quasi-identifiers in privacy-related literature. Kryszkiewicz also indicates that some objects may be *indiscernible* with regard to their description in an incomplete system though they may have different properties in reality. The notion of k-anonymity is based on the concept of k indiscernible records. In [23], a Rough Sets algorithm for computing deterministic classification rules from an incomplete information system is presented. However, Kryszkiewicz does not discuss privacy aspects of incomplete information systems.

3. Preliminaries

Here we present the terminology and notations that we shall use henceforth. We begin (Definition 3.1) by defining our novel framework of generalizations. That framework allows the generalization of the sensitive attribute too (as opposed to standard generalizations in which only the quasi-identifiers are generalized, while the sensitive values remain unchanged). In addition, the sensitive attribute is generalized in a new, probabilistic manner, by replacing each sensitive value with a frequency distribution over the sensitive domain (and not by a subset of values, as is the case with standard generalizations). Then, we formally define the closure of a set of records as the least generalized record that generalizes each of the records in the set, its information loss, and its diversity (Definition 3.2). Finally, we present our model of non-homogeneous anonymization (Definition 3.3).

Our standard assumption is that the set of possible values of each quasi-identifier is defined in the database metadata. Let A_m, $m \in [M]$, denote the set of possible values for the mth quasi-identifier, and lest A_{M+1} be the set of possible sensitive values. Let $T = \{R_1, \dots, R_N\}$ be a table of N records in $A_1 \times \cdots \times A_{M+1}$. We proceed to define our extended framework of generalizations. In that extended framework, the quasi-identifiers are generalized in the usual manner; as for the sensitive attribute, it may be generalized too, but in a different, probabilistic manner.

Definition 3.1. *Assume that:*

(a) For all $m \in [M] := \{1, \ldots, M\}$, $\overline{A}_m$ *is a given collection of subsets of* A_m*;*

(b) $\mathcal{D}(A_{M+1})$ *is the set of all frequency distributions on* A_{M+1}*, i.e., all mappings* $f : A_{M+1} \rightarrow [0,1]$ *such that* $\sum_{a \in A_{M+1}} f(a) = 1$.

Then the generalized record $\overline{R} = (\overline{R}(1), \ldots, \overline{R}(M), \overline{R}(M+1)) \in \overline{A}_1 \times \cdots \times \overline{A}_M \times \mathcal{D}(A_{M+1})$ *generalizes the record* $R \in A_1 \times \cdots \times A_M \times A_{M+1}$ *(denoted* $R \sqsubseteq \overline{R}$*) if:*

(c) For all $m \in [M]$, $R(m) \in \overline{R}(m)$*, and*

(d) $\overline{R}(M+1)(R(M+1)) > 0$*, i.e., the frequency distribution* $\overline{R}(M+1)$ *assigns a positive frequency to the original sensitive value* $R(M+1)$.

Finally, $\overline{T} = \{\overline{R}_1, \ldots, \overline{R}_N\} \subset \overline{A}_1 \times \cdots \times \overline{A}_M \times \mathcal{D}(A_{M+1})$ *is a generalization of* $T = \{R_1, \ldots, R_N\} \subset A_1 \times \cdots \times A_M \times A_{M+1}$ *if* $R_n \sqsubseteq \overline{R}_n$ *for all* $n \in [N]$.

Comments.

(i) For each quasi-identifier A_m, $m \in [M]$, $\overline{A}_m$ is a user-defined collection of subsets that are allowed to be used as generalized values. We do not make any assumption regarding $\overline{A}_m$, apart for the trivial assumption that every element $a \in A_m$ has a subset $S_a \in \overline{A}_m$ that contains it. A typical choice for $\overline{A}_m$ in the case of a categorical attribute is a taxonomy tree of A_m. For numeric attributes, $\overline{A}_m$ typically consists of all intervals. Having said that, our entire discussion herein is independent of the user's selection of those collections of subsets.

(ii) The sensitive values may be replaced with frequency distributions that support them, namely, frequency distributions that assign a positive frequency to the original value, but may "hide" it amongst other sensitive values. The customary model of generalizations is a special case of the above defined model, in which all of the frequency distributions are concentrated in the sensitive value of the original record (namely, they assign a frequency of 1 to that value, and a zero frequency to all other sensitive values).

As an example, consider Bob's record in Table 1(a). It may be generalized to

$$\overline{R}_{\text{Bob}} = (\ \text{21-55}\ ,\ \text{10***}\ ,\ \{(\text{Flu}, \frac{2}{3}), (\text{Angina}, \frac{1}{3})\}\ )\,. \tag{1}$$

The last entry in $\overline{R}_{\text{Bob}}$ is a frequency distribution over $A_3 =$`Disease` that associates with Flu the frequency 2/3 and with Angina the frequency 1/3.

Several measures of information loss were defined and used in the literature thus far. For example, the Loss Metric measure [18], which is a commonly used one, assigns the following generalization cost to a given generalized record $\overline{R}$:

$$IL(\overline{R}) = \frac{1}{M} \sum_{m=1}^{M} \frac{|\overline{R}(m)| - 1}{|A_m| - 1}\,. \tag{2}$$

Namely, this measure incurs a generalization cost for each quasi-identifier entry in the record which is proportional to the size of the subset to which it was generalized; in particular, entries that remain unchanged (namely, $|\overline{R}(m)| = 1$) will incur a cost of zero, while entries that were completely suppressed ($|\overline{R}(m)| = |A_m|$) will incur a cost of 1. Generalized tables with smaller values of the LM measure retain more information on the original values of the quasi-identifiers, whence it is plausible to expect that data mining algorithms trained on such tables will produce more accurate classification models.

Numeric attributes are expressed and stored to a finite precision. For example, `Age` can be specified by whole years, while `Weight` can be specified in kilograms up to one decimal digit after the point. Hence, the domain (or range) A_m that corresponds to numeric attributes is also finite (as is the case with categorical attributes). For numeric attributes, a typical generalized value is an interval. If $\overline{R}(m)$ is an interval, then $|\overline{R}(m)|$ denotes the number of possible values in that interval (regardless of whether all of those values appear in the data table or not). Practically, both $|\overline{R}(m)|$ and $|A_m|$ can be taken as the lengths of the corresponding intervals.

Definition 3.2. *The closure of a set of records $B \subset A_1 \times \cdots \times A_M \times A_{M+1}$ is the generalized record $\overline{B} = (\overline{B}(1), \ldots, \overline{B}(M), \overline{B}(M+1))$ where:*

(a) For all $m \in [M]$, $\overline{B}(m)$ is the minimal (with respect to inclusion) subset in $\overline{A}_m$ that includes $R(m)$ for all $R \in B$; and

(b) $\overline{B}(M+1)$ is the frequency distribution $f : A_{M+1} \to [0,1]$ that is defined by

$$f(a) = \frac{|\{R \in B : R(M+1) = a\}|}{|B|} \qquad \forall a \in A_{M+1} .$$

The information loss of B, denoted $IL(B)$, is defined as the information loss of its closure.

The diversity of B is $div(B) = \left(\max \overline{B}(M+1)\right)^{-1}$ (i.e., the inverse of the maximal frequency in the distribution $\overline{B}(M+1)$).

For example, the generalized record in Eq. (1) is the closure of Bob's, Carol's and David's records in Table 1(a). Its diversity is $\frac{3}{2}$. The information loss of the set $B = \{R_{\text{Bob}}, R_{\text{Carol}}, R_{\text{David}}\}$ that consists of the second, third and fourth records in Table 1(a), is the information loss of its closure, namely, of the generalized record $\overline{R}_{\text{Bob}}$ in Eq. (1).

Finally, we define the notion of non-homogeneous (k, ℓ)-anonymizations:

Definition 3.3. *Let $\overline{T}$ be a generalization of T in the sense of Definition 3.1. It respects non-homogeneous k-anonymity if each $\overline{R}_n \in \overline{T}$ generalizes at least k records from T. It satisfies the ℓ-diversity constraint if the maximal frequency in each of the frequency distributions $\overline{R}_n(M+1)$, $n \in [N]$, is no larger than $1/\ell$. If $\overline{T}$ is a generalization of T in the sense of Definition 3.1 that respects non-homogeneous k-anonymity and ℓ-diversity, it is called a non-homogeneous (k, ℓ)-anonymization.*

Table 2 shows a non-homogeneous ($k = 2, \ell = 2$)-anonymization of Table 1(a) with a frequency distribution generalization of the sensitive attribute. Indeed, the nth record in Table 2, $1 \leq n \leq 5$, is a generalization of the nth record in Table 1(a) and at least one more record from that table; and all the sensitive distributions include frequencies that are no larger than 1/2.

Age	Zipcode	Disease distribution
21-30	10055	$\{(\text{M}, \frac{1}{2}), (\text{F}, \frac{1}{2})\}$
21	100**	$\{(\text{F}, \frac{1}{2}), (\text{A}, \frac{1}{2})\}$
21-30	100**	$\{(\text{A}, \frac{1}{2}), (\text{M}, \frac{1}{2})\}$
47-55	10***	$\{(\text{F}, \frac{1}{2}), (\text{D}, \frac{1}{2})\}$
47-55	10***	$\{(\text{F}, \frac{1}{2}), (\text{D}, \frac{1}{2})\}$

Table 2: Non-homogeneous anonymization with sensitive value distributions

The above defined model of generalization is non-homogeneous since it does not require each generalized record to be identical to at least $k - 1$ other generalized records, when projected onto the quasi-identifiers. In that sense, it is similar to previous notions of non-homogeneous anonymizations [11, 39, 43]. However, the above defined notion differs from those in [11, 39, 43] by allowing the generalization of the sensitive attribute, and in the manner in which it defines compliance with the k-anonymity and ℓ-diversity constraints. In Section 4.5 we discuss the different notions of non-homogeneous anonymizations and compare between them.

The notion of non-homogeneous (k, ℓ)-anonymization as defined above is also related to the notion of (α, k)-anonymization [42]. Let $\overline{T}$ be a standard generalization of T, namely, a generalization as in Definition 3.1 where all sensitive frequency distributions are concentrated in the original sensitive value. Define an equivalence relation between the generalized records in $\overline{T}$ where $\overline{R}_n \sim \overline{R}_{n'}$ if $\overline{R}_n(m) = \overline{R}_{n'}(m)$ for all $m \in [M]$. Then $\overline{T}$ respects standard (or homogeneous) k-anonymity if each equivalence class in $\overline{T}/\!\sim$ is of size at least k. It respects ℓ-diversity if the relative frequency of each sensitive value within each equivalence class is no larger than $1/\ell$. Finally, it respects $(1/\ell, k)$-anonymity [42] if it respects homogeneous k-anonymity as well as ℓ-diversity. The above defined notion of non-homogeneous (k, ℓ)-anonymity differs from $(1/\ell, k)$-anonymity in two aspects: (a) It includes non-homogeneous anonymizations (rather than only homogeneous ones); and (b) it allows publishing sensitive value distributions in each generalized record, and enforces ℓ-diversity through those distributions (rather than publishing a single sensitive value for each record, and then enforcing ℓ-diversity through the sensitive value distribution within each equivalence class).

4. An algorithm for Non-homogeneous generalization with Sensitive Value Distribution (NSVDist)

4.1. The algorithm

Our Non-homogeneous generalization algorithm with Sensitive Value Distribution (NSVDist) (Algorithm 1) produces for each record $R_n \in T$ a corresponding generalized record $\overline{R}_n$ which is the closure of R_n and $k-1$ additional records in T; the subset of T that includes R_n and the additional $k-1$ records is denoted B_n. The selection of the $k-1$ additional records in B_n is guided by two rules — one that relates to the generalized quasi-identifiers and another that relates to the sensitive distribution: (a) Trying to minimize the resulting information loss IL of B_n; and (b) making sure that the diversity of B_n is at least ℓ. The selection is carried out in a greedy manner: The $k-1$ records that will be used to mask a given record $R_n \in T$ are selected one at a time, where in each stage we select a record that complies with the diversity constraint and minimizes the resulting information loss due to generalization. The operation of the algorithm is independent of the choice of information loss measure. (In our experiments we implemented it with the LM measure, Eq. (2), and the entropy measure [12].)

To that end, in order to compute the generalization $\overline{R}_n$ of the record R_n, $n \in [N]$, we compute a set B_n that includes R_n and additional $k-1$ records, so that the diversity of B_n is at least ℓ and its information loss is as small as possible (Lines 2-7). The set B_n is initialized to include only R_n (Line 2). Then we start adding to it one additional record at a time until its size becomes k (Lines 4-7). In order to verify the diversity constraint, we maintain a frequency vector F of length $|A_{M+1}|$ (the number of sensitive values) so that at each stage $F(q)$ equals the number of records in B_n whose sensitive value is the qth value in A_{M+1}. That vector is initialized in Line 3 for the initial set B_n. In the loop that implements the greedy selection, we review all records that were not selected yet. We skip records that cannot be added to B_n without violating the diversity constraint. Specifically, since B_n will eventually be of size k, it will be ℓ-diverse if and only if it does not contain more than $\lfloor k/\ell \rfloor$ records that have the same sensitive value. Hence, we concentrate only on records whose sensitive value appears in B_n strictly less than $\lfloor k/\ell \rfloor$ times. Among all those records, we select the one, R_i, whose addition to B_n would yield a set $B_n \cup \{R_i\}$ of minimal information loss (Line 5). The function that computes the information loss of a given set of records is described in Algorithm 2. After selecting that record, we add it to B_n and update the vector F accordingly (Line 6). At the end, when B_n includes R_n and additional $k-1$ records, we set $\overline{R}_n$ to be the closure of B_n. That computation is also described in Algorithm 2. (Since Algorithm 2 is a straightforward implementation of Definition 3.2 and Eq. (2), it is self-explanatory.)

Each generalized record in the output anonymization $\overline{T}$ is consistent with R_n and at least $k-1$ other records in T. In addition, since none of the sensitive values in A_{M+1} appears in more than k/ℓ of those k records, the frequency of each sensitive value in each of the frequency distributions in $\overline{T}$ is no more than $1/\ell$. Therefore, the output of Algorithm 1 is a (k,ℓ)-anonymization of T.

Algorithm 1 Non-homogeneous generalization with Sensitive Value Distribution (NSVDist)

```
Input: A table $T = \{R_1, \ldots, R_N\}$, anonymity parameter $k$, diversity parameter $\ell$.
Output: A non-homogeneous $(k,\ell)$-anonymization $\overline{T} = \{\overline{R}_1, \ldots, \overline{R}_N\}$ with sensitive value
    distribution.
for all $1 \leq n \leq N$ do
    Set $B_n = \{R_n\}$.
    Set $F(q) = 0$ for all $q \in A_{M+1} \setminus \{R_n(M+1)\}$ and $F(q) = 1$ for $q = R_n(M+1)$.
    while $|B_n| < k$ do
        Among all records $R_i \in T \setminus B_n$ for which $F(R_i(M+1)) < \lfloor k/\ell \rfloor$, find one that
            minimizes $IL(B_n \cup \{R_i\})$. {See Algorithm 2.}
        Add the selected $R_i$ to $B_n$ and set $F(R_i(M+1)) = F(R_i(M+1)) + 1$.
    end while
    $\overline{R}_n = \overline{B_n}$. {See Algorithm 2.}
end for
Return $\overline{T} = \{\overline{R}_1, \ldots, \overline{R}_N\}$.
```

4.2. Information loss

In our first set of experiments, we compared the information loss in non-homogeneous (k,ℓ)-anonymizations, as issued by the NSVDist algorithm, to the information loss in corresponding (k,ℓ)-anonymizations, as issued by leading homogeneous anonymization algorithms. To that end, we used the single-dimensional Mondrian algorithm [25] and the sequential anonymization algorithm [13] (SeqA)[1]. We conducted this set of experiments on three datasets from the UCI Machine Learning Repository [6] — `Adult`, `CMC`, and `Mammographic`. (Our experimental setup is described in more detail in Section 5.2.) Figure 1 shows the average information losses, as measured by the LM measure, in the output anonymizations of NSVDist, SeqA, and Mondrian on these three datasets, for various values of k and two values of the diversity parameter ℓ. (Specifically, we measured the information loss in each of the generalized records in the output anonymization, using Eq. (2), and then divided by the number of records.) As can be seen, non-homogeneous anonymizations yield information losses that are considerably smaller than homogeneous anonymizations. We repeated the experiments with the entropy measure of [12], instead of the LM; the results were consistent with those shown in Figure 1 for the LM measure.

Thus, our research hypothesis is that data mining algorithms trained on tables anonymized by NSVDist would produce more useful (e.g, more accurate) classification models. This

[1] We used here and later on a modified version of the sequential anonymization algorithm. While the original algorithm in [13] began with a random partition of the dataset records, and then sequentially improved that partition until reaching a local optimum, we began with the partition that was issued by the single-dimensional Mondrian algorithm. This modified version performed better than the original sequential algorithm.

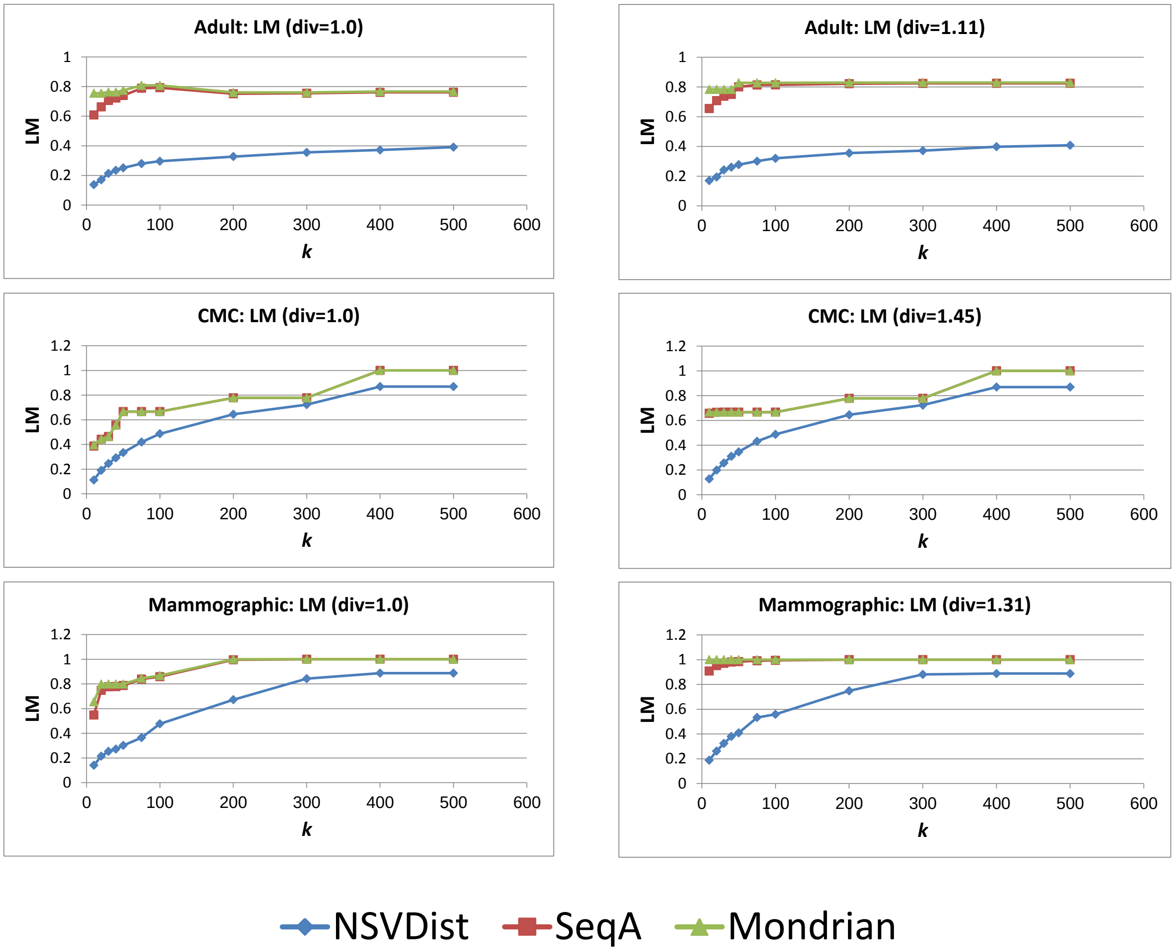


Figure 1: Loss Measure as a function of k

Algorithm 2 Computing the closure and information loss of a set of records

```
Input: A set $B = \{S_1, \dots, S_H\}$ of $H$ records from $T = \{R_1, \dots, R_N\}$.
Output: The closure $\overline{B} = (\overline{B}(1), \dots, \overline{B}(M), \overline{B}(M+1))$ and information loss $IL(B)$ of $B$
    (Definition 3.2)
$IL(B) = 0$.
for all $1 \leq m \leq M$ do
    Set $\overline{B}(m)$ to be the minimal subset in $\overline{A}_m$ that includes $S_h(m)$ for all $1 \leq h \leq H$.
    $IL(B) = IL(B) + \frac{|\overline{B}(m)|-1}{|A_m|-1}$
end for
$IL(B) = IL(B)/M$
$F = (0, \dots, 0)$ {A vector of length $s := |A_{M+1}|$, the number of sensitive values.}
for all $1 \leq h \leq H$ do
    $a = S_h(M+1)$
    $F(a) = F(a) + 1/H$
end for
$\overline{B}(M+1) = F$
Return $IL(B)$ and $\overline{B} = (\overline{B}(1), \dots, \overline{B}(M), \overline{B}(M+1))$.
```

hypothesis is examined in Section 5.

4.3. Computational complexity

The computational complexity of NSVDist is $O(kN^2M)$, since the number of candidate records that needs to be checked in Step 5 is $O(N)$, and those searches are repeated $k-1$ times for each of the N records; the linear dependence on M is due to the computation of the information loss. In case the table is too large to allow such a runtime, we may first apply on T a small number of steps of a top-down clustering algorithm that is guided by the information loss similarity measure; such a preprocessing step will split the N table records into smaller clusters of records that are close with respect to the information loss measure. After doing so, we may proceed to apply NSVDist within each cluster separately. One natural choice for such a top-down clustering algorithm is the Mondrian algorithm [25]. Splitting the table to p clusters that have similar number of records will reduce the runtime by a factor of p to $O(pk(N/p)^2M) = O(kN^2M/p)$.

4.4. Privacy

NSVDist produces (k, ℓ)-anonymizations that provide the same level of privacy as do homogeneous k-anonymizations that are ℓ-diverse. Both types of anonymization link each target record with a *multiset*[2] of sensitive values. In both cases, that multiset contains the true sensitive value of the target record together with the sensitive values of at least $k-1$ additional records. For example, the generalized record $\overline{R}_{\text{Bob}}$ in Eq. (1) is a generalization

[2] A multiset is a set with possibly repeating elements.

of Bob's record in Table 1(a). It is the closure of Bob's, Carol's and David's records. The frequency distribution in the sensitive column is equivalent to the multiset of sensitive values $\{\mathrm{Flu}, \mathrm{Flu}, \mathrm{Angina}\}$ which holds the sensitive value of Bob as well as those of Carol and David. If, on the other hand, we consider the homogeneous anonymization of Table 1(a) in Table 1(b), it links Bob's record to the multiset $\{\mathrm{Measles}, \mathrm{Flu}, \mathrm{Angina}\}$, that holds the sensitive values of Alice, Bob, and Carol.

Both non-homogeneous (k, ℓ)-anonymizations and homogeneous k-anonymizations that are ℓ-diverse require that none of the sensitive values in those multisets appear in frequency that is greater than $1/\ell$. Assume that an adversary will attempt to gain knowledge on the sensitive values of some of the individuals behind the masking values in the multiset of his target record, in order to learn more information on the sensitive value of his target record. Adopting that strategy, he will be able to infer the sensitive value of his target record with certainty once he gains knowledge of the sensitive values of all individuals whose sensitive value differs from that of his target record. By ℓ-diversity, there are at least $\lceil k(1 - \ell^{-1}) \rceil$ such individuals. Hence, the combination of the k-anonymity and ℓ-diversity conditions imply the same lower bound on the number of individuals for which the adversary needs to learn the sensitive information in both anonymization models.

4.5. Comparison to other non-homogeneous anonymization algorithms

Non-homogeneous anonymization was introduced in [11] and then further explored in [39] and [43]. Both studies presented algorithms for computing non-homogeneous anonymized views and compared their performance against homogeneous anonymization algorithms in terms of general purpose information loss measures; the former study used the LM measure, Eq. (2), and the entropy measure of [12], while the latter used only the LM measure.

Gionis et al. [11, 39] concentrated on achieving a privacy goal that "simulates" k-anonymity. One algorithm presented there computed a non-homogeneous anonymization $\overline{T}$ of T that has the following property: Every record $R \in T$ is consistent with at least k generalized records in $\overline{T}$ and, on the other hand, every generalized record in $\overline{T}$ is consistent with at least k records in T. The two tables T and $\overline{T}$ induce a bipartite graph in which an edge connects $R \in T$ with $\overline{R} \in \overline{T}$ if and only if $R \sqsubseteq \overline{R}$. Hence, stated otherwise, the output of that algorithm is a non-homogeneous anonymization $\overline{T}$ of T for which the degrees of all nodes in the resulting bipartite graph are at least k. It was argued there that such anonymizations provide in practice anonymity that is comparable to k-anonymity. However, if the adversary is assumed to know all quasi-identifiers of all records in T, he may be able to reproduce the entire bipartite graph and then ignore edges that are not part of a perfect matching, since such edges cannot stand for true links between records and their generalized image. By doing so, the effective degrees of nodes may become smaller than k. Their second algorithm addresses that privacy breach by making sure that each node $R \in T$ has at least k nodes $\overline{R} \in \overline{T}$ that are connected to R by an edge which is included in a perfect matching.

That work did not consider ℓ-diversity. Hence, such non-homogeneous anonymizations may leak sensitive information in the same way that k-anonymizations that do not respect

ℓ-diversity might.

The algorithm proposed by Wong et al. [43] rectifies that problem. It starts by applying a top-down clustering algorithm in order to split the records into clusters such that the diversity within each cluster is at least ℓ, on one hand, and the records within each cluster are "close" in terms of the underlying information loss measure, on the other hand. The rest of the algorithm proceeds within each cluster independently. Their algorithm, like NSVDist, generates for each record $R_n \in T$ a generalized view $\overline{R}_n$ that is the closure of R_n and a number of other records ($\ell - 1$ other records in their algorithm). Then, they attach to $\overline{R}_n$ one of the sensitive values that belong to one of the original records that were generalized by it. The selection of that sensitive value is made at random so that even an adversary who knows all quasi-identifiers in T and also knows the anonymization algorithm cannot link any sensitive value to any record in T with probability greater than $1/\ell$.

We identify two limitations of this approach. The first one relates to privacy: The algorithm of [43] outputs anonymizations that are ℓ-diverse only; however, as we proceed to explain, ℓ-diversity must be enforced on top of k-anonymity, in order to get (k, ℓ)-anonymity (Definition 3.3), since it is insufficient by itself. The diversity of any anonymization of a table is bounded by the diversity of the entire table, and the latter is bounded by the number of possible sensitive values. Therefore, if the table has a sensitive attribute with a small number of possible values, all of its anonymizations will respect ℓ-diversity with ℓ that does not exceed that number. For example, in the case of a binary sensitive attribute, one can aim at achieving ℓ-diverse anonymizations with $\ell \leq 2$ only. In such a case, if one imposes only ℓ-diversity, the blocks of indistinguishable records could be of size 2. Such small blocks do not provide enough privacy for the individuals in them, because if an adversary may be able to learn the sensitive value of one of those individuals, he may infer that of the other one as well. If, on the other hand, we demand that such ℓ-diverse anonymizations are also k-anonymous, for some larger value of k, then the adversary would have to find out the sensitive values of at least $k/2$ individuals before he would be able to infer the sensitive value of his target individual. Indeed, NSVDist is designed to achieve (k, ℓ)-anonymizations in order to achieve such enhanced security. (As mentioned in Section 2, LKC-privacy [32] and (α, k)-anonymity [42] also combine the k-anonymity and ℓ-diversity conditions, but within the framework of homogeneous anonymizations.)

The second limitation of the algorithm of [43] is that it can work only with integer values of ℓ. Restricting the diversity parameter ℓ to integer values limits the applicability of the algorithm. For example, in the `Adult` dataset from the UCI Machine Learning Repository [6], which frequently serves as a benchmark dataset in this context, the sensitive value is binary and the global diversity is 1.33 (namely, the more frequent sensitive value has a frequency of $1/1.33 \approx 0.752$ in the table). In such cases, it is impossible to apply the algorithm of [43] with $\ell > 1$; indeed, the algorithm starts with ordering all table records so that each ℓ consecutive records have ℓ different sensitive values, and such an ordering does not exist for the `Adult` dataset for any integer $\ell > 1$. Similar problems will also occur with richer sensitive attribute domains where the global diversity is low. In contrast, NSVDist enforces diversity in a way that can be applied with any real value of ℓ.

To summarize, NSVDist enhances the algorithms of [11, 39, 43] by offering non-homogeneous anonymizations that are both k-anonymous and ℓ-diverse (Definition 3.3); in addition, it enhances the algorithm of [43] by allowing non-integer values of ℓ. These enhancements of the anonymization framework have been enabled by allowing the sensitive attributes to be generalized to sensitive value distributions rather than exact values.

5. Evaluation

5.1. Evaluation methodology

We evaluated the proposed anonymization methodology on several benchmark datasets using different classification algorithms. For each dataset and each classification algorithm, we carried out an evaluation procedure that consisted of the following steps:

(1) If the dataset had no available training-test partition, we applied on it the 10-fold cross-validation using the Split Data operator in the RapidMiner software (version 5.1.001) [30]. In our work, the training set is used to generate the published anonymized data that may be accessible by anyone to induce a classification model; the test set, on the other hand, represents data that is unknown at the time of performing the anonymization, and it is available only to the user of the classification model.

(2) We performed (k, ℓ)-anonymizations of the training set, for various settings of k and ℓ, using four algorithms: The sequential anonymization algorithm of [13], the single-dimensional Mondrian algorithm [25], the privacy-aware information sharing algorithm (PAIS) of [32] and NSVDist (see Section 4.1).

(3) For every setting of k and ℓ, we trained a classifier on each of the four resulting anonymized tables, using different classification algorithms. We then computed the classifier's predictive performance on the test records.

(4) In addition, we computed the accuracy of a classifier based on the majority rule, i.e., a classifier which assigns the majority class in the training set to each record in the test set. Such a classifier provides the maximum possible level of privacy, since instead of publishing the database it only publishes the majority class of the sensitive attribute. We also computed the accuracy of a classifier that was trained on the original training records (without applying anonymization of any kind); this corresponds to setting $k = \ell = 1$. Those two classifiers served as baselines in the comparison.

The standard classification algorithms cannot be applied directly on generalized tables since they contain non-specific values such as numeric intervals or subsets of nominal values. Hence, it is needed first to convert the anonymized tables into tables with specific values and only then to apply the classification algorithm on those non-generalized tables. We converted the anonymized tables into non-generalized ones by means of sampling. Assume that $\overline{R} = (\overline{R}(1), \ldots, \overline{R}(M), \overline{R}(M+1))$ is a generalized record in an anonymized table that was produced by one of the anonymization algorithms. For all quasi-identifier attributes

$m \in [M]$, $\overline{R}(m)$ is a value from $\overline{A}_m$, namely, it is a subset of the mth quasi-identifier domain A_m. As for the sensitive attribute $\overline{R}(M+1)$, it is published as a distribution over the sensitive attribute domain A_{M+1}. (In tables produced by the standard homogeneous anonymization algorithms, namely, the sequential anonymization, the Mondrian, and the PAIS algorithms, the sensitive value may be viewed as a deterministic distribution since these algorithms keep the sensitive values unchanged.) Then, we sample from $\overline{R}$ a specific record $R = (R(1), \ldots, R(M), R(M+1)) \in A_1 \times \cdots \times A_M \times A_{M+1}$ in the following manner:

- For all $m \in [M]$, $R(m)$ is one of the values in the subset $\overline{R}(m) \subseteq A_m$, drawn from the distribution of the values in the subset $\overline{R}(m)$ in the entire training set. (Those distributions can be published.) Specifically, if $\overline{R}(m)$ is a subset that includes q values from A_m, and their frequencies in the training set are $f_1, \ldots, f_q$, then we select the ith value with probability $f_i / \sum_{j=1}^{q} f_j$.

- $R(M+1)$ is a value in A_{M+1} that is drawn at random, where $\text{Prob}(R(M+1) = a) = \overline{R}(M+1)(a)$ for all $a \in A_{M+1}$.

For each of the anonymization algorithms, we repeated the sampling procedure $p = 10$ times in order to induce p models with each classifier. We report the average accuracy over those p independent samples and over the 10 training-test partitions of the 10-fold cross-validation methodology. Namely, the reported average accuracy is over $10p = 100$ classifiers.

The results presented here are for anonymizations that were issued by NSVDist, Mondrian, and the sequential anonymization algorithms, when they used the LM information loss measure. Anonymizations that were computed by those algorithms using the entropy information loss measure exhibited very similar behavior. As for the PAIS algorithm, it uses the InfoGain utility measure, which is designed for maximizing classification accuracy.

5.2. Experimental setup

We conducted our experiments on eight datasets from the UCI Machine Learning Repository [6]. Table 3 provides information on the number of records in each dataset (indicating in the parentheses the number of records removed due to a missing attribute value or a missing label), the number and the list of statistically relevant quasi-identifiers, and the global diversity. Out of the eight datasets that we used for our evaluation, only the `Adult` dataset has a given training-test partition; hence, in that dataset we did not apply the 10-fold cross validation methodology and, consequently, the accuracy values reported for that dataset are an average over the p independent samples.

The statistically relevant quasi-identifiers were detected by applying on each dataset the Weka software (version 3.68) [14] operator ”CfsSubsetEval” with ”BestFirst” search method for attribute selection. This method, which is based on greedy hill-climbing and backtracking search, chooses a subset of attributes having the highest predictive value, along with a low degree of redundancy among them.

As for the global diversity, it equals the inverse of the maximal frequency of a sensitive value in the whole dataset. For example, a diversity of 1.94 in the `Mammographic Mass` dataset indicates that the most frequent sensitive value appears in 51.54% of the records. In each dataset, the diversity of a given generalization cannot exceed the global diversity.

Dataset	Records	Quasi-identifiers	Diversity
`Abalone`	4177 (0)	5: sex, diameter, height, viscera_ weight, shell_ weights	6.06
`Adult`	45222 (3620)	14: age, work_class, final_weight, education, education_num, marital_status, occupation, relationship, race, sex, capital_gain, capital_loss, hours_per_week, native_country	1.33
`Breast Cancer Wisconsin (Original)`	683 (16)	6: ct, uocsi, uocsh, bn, bc, nn	1.54
`Contraceptive Method Choice`	1473 (0)	3: wife_age, wife_edu, num_children	2.34
`Ecoli`	336 (0)	6: seq, mcg, gvh, lip, alm1, alm2	2.35
`Mammographic Mass`	830 (131)	5: bi_rads_assessment, age, shape, margin, density	1.94
`Page Blocks Classification`	5473 (0)	6: height, eccen, p_black, p_and, mean_tr, wb_trans	1.11
`Yeast`	1484 (0)	4: seq, alm, erl, pox	3.21

Table 3: Datasets

For the `Adult` dataset, we used the taxonomy trees that were suggested by Mohammed et. al [32]. In all other datasets, we built artificial taxonomy trees for all attributes using the following automatic procedure. We sorted the set S of possible values in the attribute alphabetically, and then constructed a tree of height $\lceil \log_5 |S| \rceil$, where each node has at most 5 children. In such a tree, the root represents the entire set S, the leaves are all singleton values, and each intermediate node represents the union of the subsets represented by its children. The obtained trees are available from the authors upon request.

In each dataset, we used several k values, starting from $k = 1$ (which corresponds to the non-generalized training dataset) and then continuing with larger values of k until we reached total suppression with most anonymization algorithms. We also tested two values of the diversity parameter ℓ: $\ell = 1$ and $\ell = 1 + (\ell_g - 1)/3$, where ℓ_g is the global diversity of the entire dataset as reported in Table 3. We repeated our experiments with four popular classification algorithms — W-J48 (based on C4.5 [34]), Naïve Bayes [15], W-JRip [4], and SVM [16] using the Weka software (version 3.68) [14].

• For W-J48 we used the following default settings: use unpruned tree (U)=false; Confidence threshold for pruning (C)=0.25; minimum number of instances per leaf (M)=2; use reduced error pruning (R)=false; number of folds for reduced error pruning(N)=3; use binary splits only (B)=false; do not perform subtree raising (S)=false; do not clean up after the tree has been built (L)=false; Laplace smoothing for predicted probabilities (A)=false; seed for random data shuffling (Q)=1.

- For Naïve Bayes (Kernel) we used the following settings: use Laplace correction to prevent high influence of zero probabilities = true; the kernel density estimation mode = full; the method to set the kernel bandwidth = heuristic; use a kernel density function grid in model application = false.

- For W-JRip we used the following default settings: The number of folds for reduced error pruning(F)=3; one fold is used as the pruning set; the minimal weights of instances within a split (N)=2.0; the number of runs of optimizations (O)=2; turn on the debug mode (D)=false; the seed of randomization (S)=1; not check the error rate ≥ 0.5 in stopping criteria (E)=false; not use pruning (P)=false.

- For SVM we used the following default settings: SVM for classification=C-SVC; the type of the kernel functions=rbf; the parameter gamma=0; the cost parameter C=0; the cache size in Megabyte=80; the tolerance of termination criterion (ϵ)=0.001; use the shrinking heuristics=true; calculate confidence values=false; select the class with the highest confidence in the multiclass setting=true.

Regarding the anonymization algorithms, we have implemented all algorithms, except for PAIS. The software for the latter algorithm was provided to us by Dr. Benjamin C. M. Fung.[3]

5.3. Experimental results

Figures 2—4 herein, and Figures B.7—B.11 in Appendix B, show the trade-off between the anonymity level k of the training data and the testing accuracy of the four evaluated classifiers, in each of the eight datasets. The left column of plots in each figure show the classifier accuracy when the training data was anonymized with the diversity parameter $\ell = 1$ (namely, in anonymizations with a trivial diversity constraint), while the right column of plots show the results with a higher diversity parameter, as explained earlier. Each plot in each of those figures includes four curves, representing the sequential anonymization algorithm (SeqA), the Mondrian algorithm, the privacy-aware information sharing algorithm (PAIS), and the NSVDist algorithm. In addition, each plot includes two reference baselines — the classification accuracy based on the majority rule and the accuracy of a classifier that was trained on the original data records. Each point on every curve represents the average over 10 independent samples (of a specific table from the anonymized table with generalized values) and over 10 training-test partitions, whenever the dataset had no training-test partition.

Table 4 provides another succinct look at the results of the above described series of

[3] The PAIS algorithm assumes that the database has two different attributes — class attribute and sensitive attribute. Since in our datasets the sensitive attribute is identical to the class attribute (as assumed by most anonymization algorithms), we created, for the sake of PAIS, a new class attribute that coincides with the sensitive attribute.

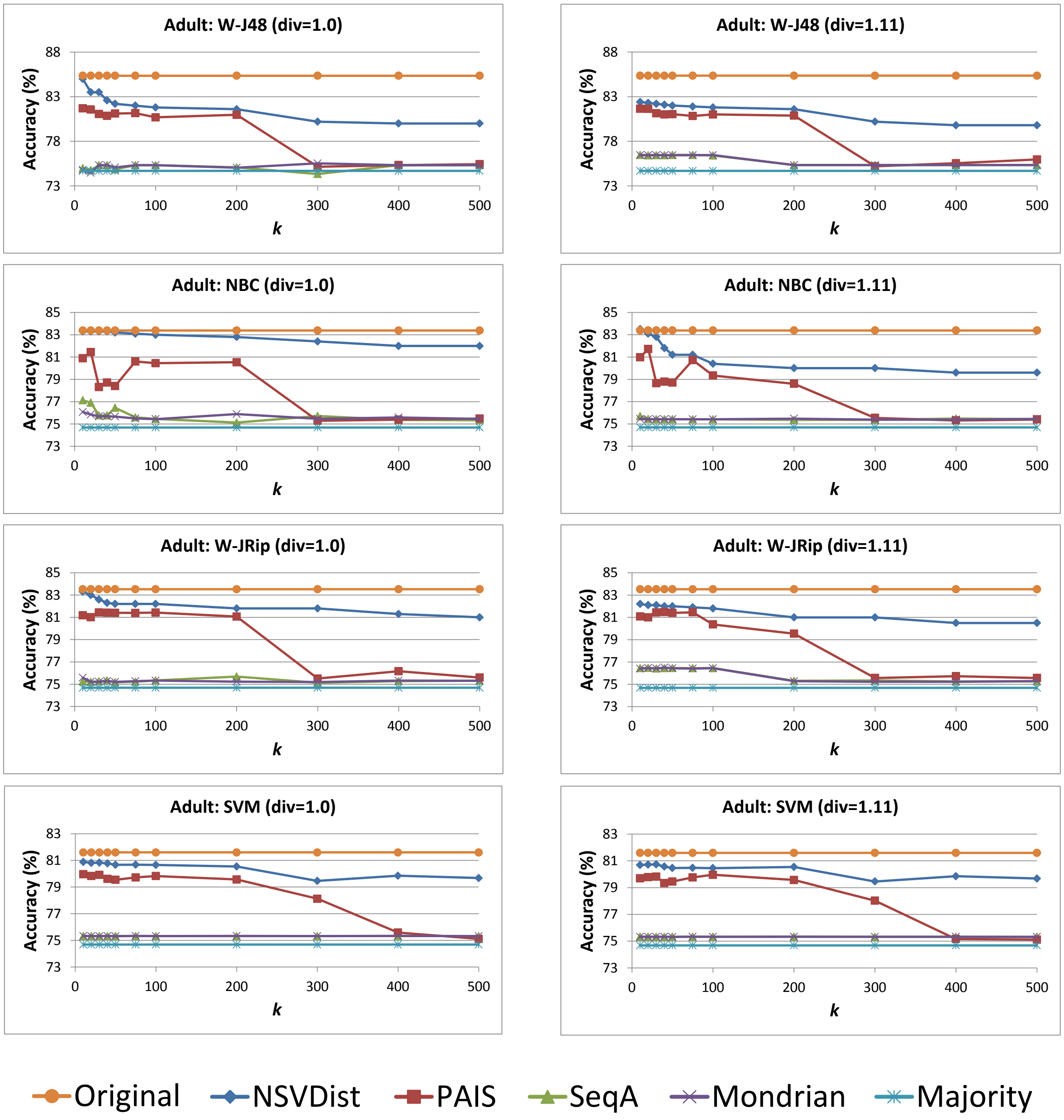


Figure 2: Classification performance using anonymized data (Adult)

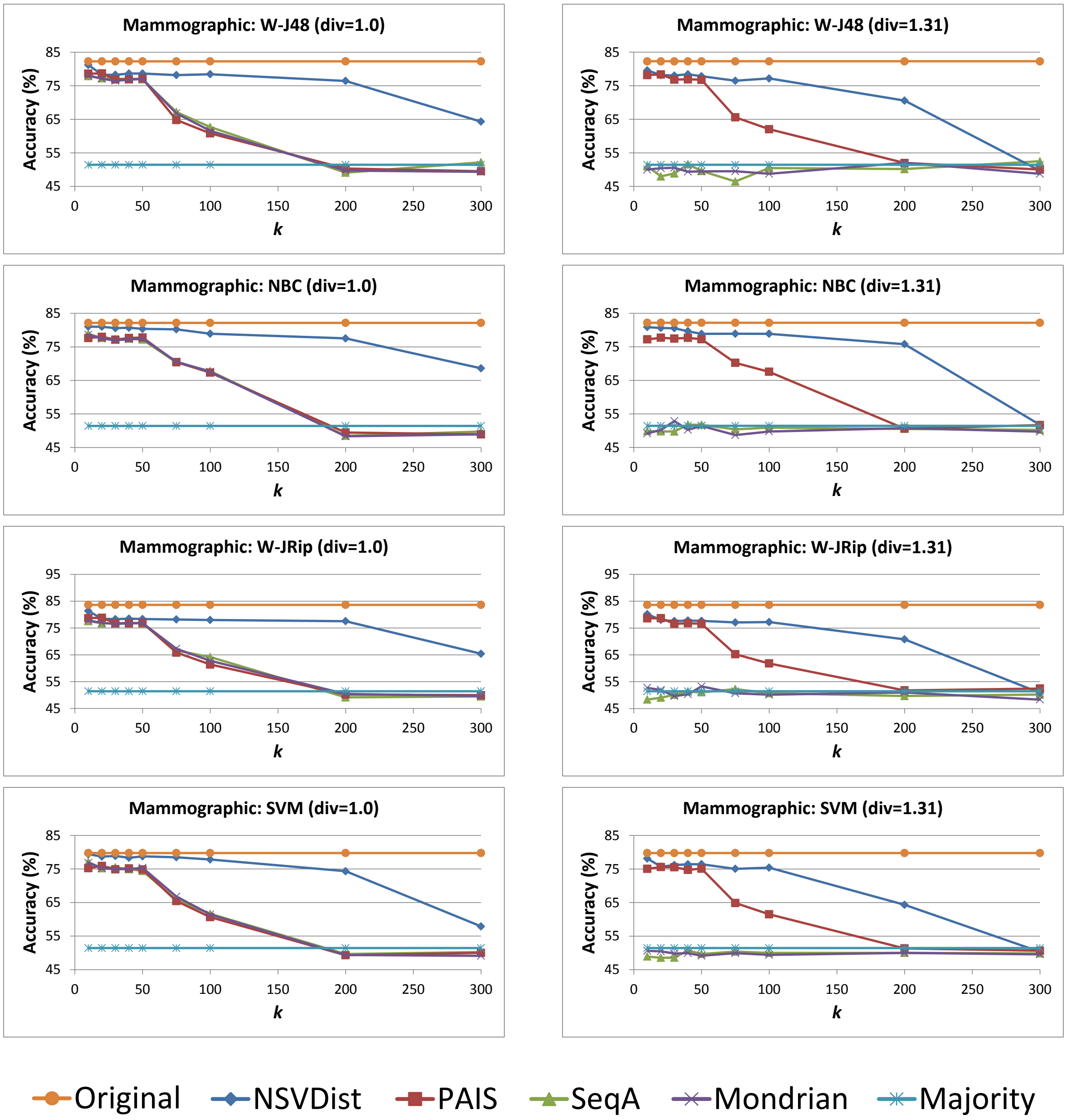


Figure 3: Classification performance using anonymized data (Mammographic)

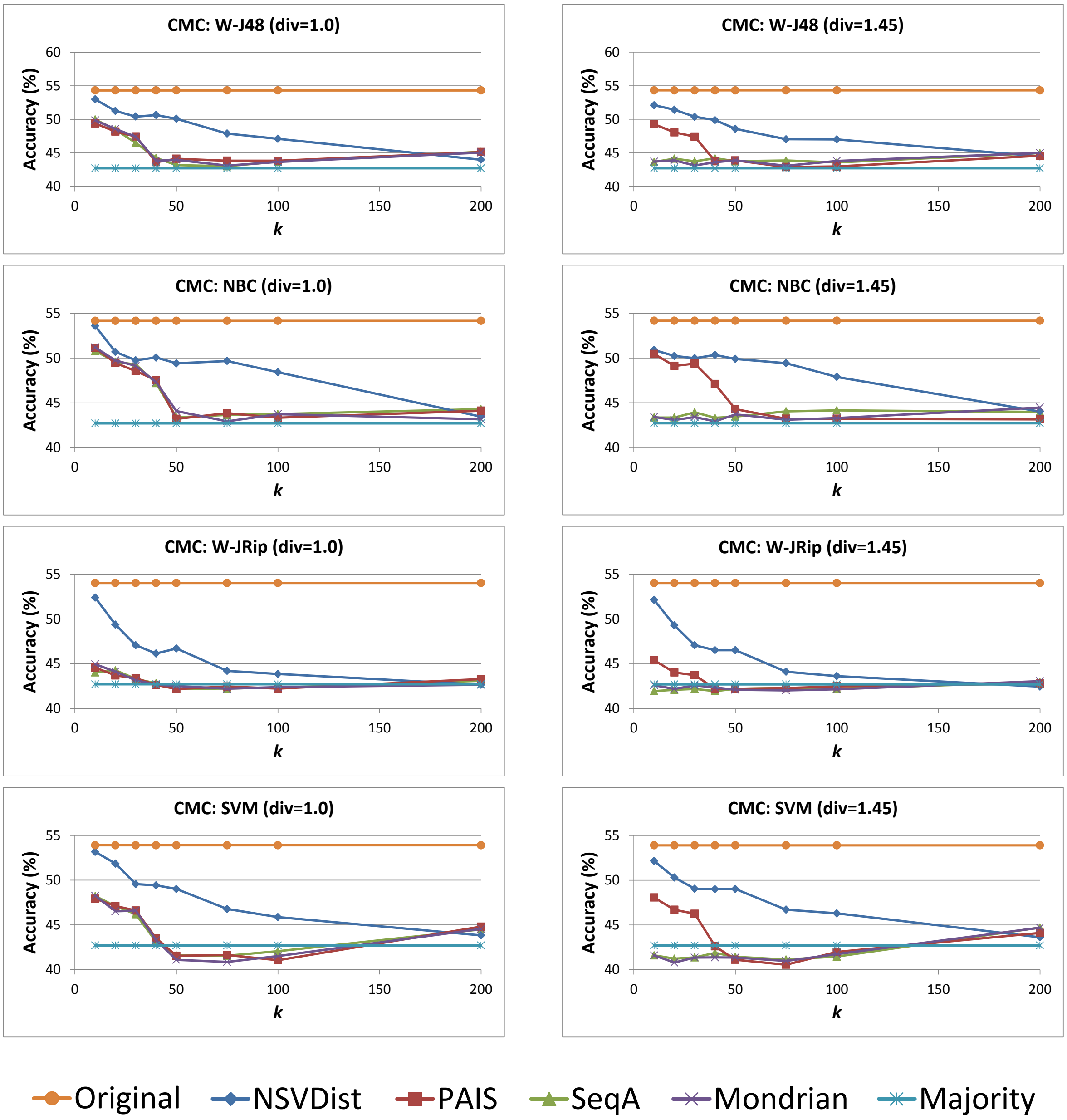


Figure 4: Classification performance using anonymized data (CMC)

Dataset	div=1.0				div>1.0				Majority	Original
	NSVDist	PAIS	SeqA	Mondrian	NSVDist	PAIS	SeqA	Mondrian		
Abalone	**19.20**	13.69	13.92	13.87	**19.24**	13.69	13.78	13.59	16.50	20.25
Adult	**82.20**	81.11	74.87	75.07	**82.00**	81.04	76.47	76.44	74.69	85.35
Breast Cancer	**91.13**	84.63	82.44	82.72	83.49	**84.76**	68.25	68.34	65.01	95.16
CMC	**50.08**	44.13	43.18	43.97	**48.58**	43.87	43.74	43.88	42.70	54.31
Ecoli	**61.65**	42.06	42.06	42.06	**51.24**	42.06	41.97	42.06	42.56	78.53
Mammographic	**78.65**	77.14	76.92	77.01	**77.84**	76.77	49.58	49.47	51.45	82.29
Page Blocks	**91.14**	90.24	89.76	89.72	**91.45**	90.47	89.76	89.74	89.77	97.06
Yeast	**39.06**	30.85	30.96	31.01	**38.97**	31.03	30.86	31.10	31.20	41.11

Dataset	div=1.0				div>1.0				Majority	Original
	NSVDist	PAIS	SeqA	Mondrian	NSVDist	PAIS	SeqA	Mondrian		
Abalone	**22.07**	19.35	19.30	18.88	**22.11**	19.48	19.56	19.31	16.50	25.23
Adult	**83.20**	78.41	76.46	75.67	**81.20**	78.71	75.42	75.43	74.69	83.38
Breast Cancer	**95.85**	86.12	85.63	85.12	**86.44**	86.37	69.96	72.18	65.01	96.48
CMC	**49.41**	43.20	43.37	44.08	**49.90**	44.29	43.46	43.71	42.70	54.17
Ecoli	**62.06**	39.94	37.74	37.97	**59.09**	38.38	38.82	39.56	42.56	80.66
Mammographic	**80.36**	77.78	77.18	77.48	**78.86**	77.27	51.73	51.42	51.45	82.17
Page Blocks	**86.65**	81.90	67.16	68.84	**81.99**	79.30	69.70	72.73	89.77	94.41
Yeast	**38.98**	29.43	29.36	30.20	**38.86**	29.03	29.76	29.39	31.20	38.88

Dataset	div=1.0				div>1.0				Majority	Original
	NSVDist	PAIS	SeqA	Mondrian	NSVDist	PAIS	SeqA	Mondrian		
Abalone	**17.90**	16.49	16.52	16.57	**17.91**	16.49	16.55	16.52	16.50	18.60
Adult	**82.20**	81.40	75.17	75.21	**82.00**	81.40	76.46	76.44	74.69	83.52
Breast Cancer	**92.56**	84.21	81.79	82.53	**88.34**	84.34	68.04	67.72	65.01	94.58
CMC	**46.71**	42.17	42.16	42.52	**46.53**	42.21	42.20	42.10	42.70	54.04
Ecoli	**52.06**	41.74	41.15	41.29	**48.62**	41.59	42.00	41.53	42.56	78.00
Mammographic	**78.33**	76.94	76.67	76.83	**77.67**	76.61	51.05	53.25	51.45	83.61
Page Blocks	**91.37**	90.54	89.69	89.67	**91.57**	90.59	89.71	89.69	89.77	96.91
Yeast	**32.66**	30.98	31.01	31.01	**32.24**	31.00	31.06	31.01	31.20	39.69

Dataset	div=1.0				div>1.0				Majority	Original
	NSVDist	PAIS	SeqA	Mondrian	NSVDist	PAIS	SeqA	Mondrian		
Abalone	**22.71**	19.29	19.26	19.31	**22.37**	19.26	19.31	19.28	16.50	23.25
Adult	**80.66**	79.55	75.33	75.32	**80.46**	79.46	75.33	75.32	74.69	81.60
Breast Cancer	**95.04**	66.78	65.13	65.93	65.04	**67.21**	58.51	58.24	65.01	96.34
CMC	**49.02**	41.54	41.59	41.09	**49.02**	41.11	41.44	41.35	42.70	53.90
Ecoli	**37.29**	34.44	33.82	33.41	**36.06**	34.12	33.97	35.18	42.56	85.43
Mammographic	**78.78**	74.88	74.40	75.37	**76.43**	75.11	49.69	49.16	51.45	79.76
Page Blocks	89.72	**89.76**	89.74	89.75	89.74	**89.75**	89.74	89.75	89.77	92.05
Yeast	**29.66**	27.59	27.20	27.72	**29.00**	26.89	27.43	27.60	31.20	93.20

Table 4: Accuracy at $k = 50$: W-J48 (top), Naïve Bayes, W-JRip, and SVM (bottom)

experiments. Each of the four tables in it reports the results for one of the four classification algorithms — W-J48, Naïve Bayes, W-JRip, and SVM. In each table, there are eight rows corresponding to the eight data sets, and ten columns: four columns that show the accuracy of a classifier that was trained on the anonymized training data with a representative anonymity parameter $k = 50$ and diversity $\ell = 1$ (one column for each of the anonymization algorithms); the next four columns give the accuracy when the diversity parameter was set to a higher value, as we explained earlier; and the last two columns give the two baseline values. In each row, the best value among the results with $\ell = 1$ (the first group of four columns) is highlighted and so is the best value among the results with $\ell > 1$ (the second group of four columns).

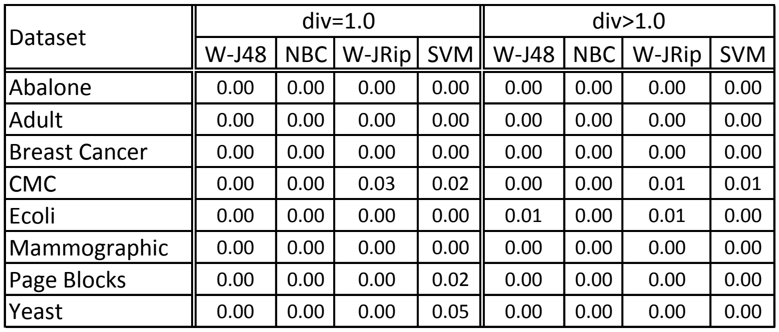
As can be seen in Table 4, the classifier that was trained on NSVDist-anonymized data was almost always the most accurate one. There were only 4 exceptions (out of 64 times) in which the PAIS-based classifier was better than the NSVDist-based classifier. Recall that PAIS is an anonymization algorithm that is targeted towards maximizing the utility for classification, while NSVDist is not targeted towards a specific data mining task.

Figure 2 shows the results with the `Adult` dataset. Here, for all values of k and ℓ, the NSVDist-classifier was more accurate than the other three classifiers; PAIS was almost always the second best. While the accuracy of the sequential- and Mondrian-based classifiers collapsed to the majority baseline for $k \geq 200$ and that of the PAIS-based classifier for $k \geq 300$, the NSVDist-based classifier continued to produce meaningful accuracy up to $k = 800$ (we show here the results only up to $k = 500$). Hence, NSVDist-anonymizations can double and even triple the level of anonymity, compared to the other algorithms, and still provide better utility. Similar behavior occurs with the `Mammographic` (Figure 3) and `CMC` datasets (Figure 4), but here the collapse of the NSVDist-classifier to the majority baseline occurs for smaller values of k, since those datasets are smaller than `Adult`.

In summary, almost all models based on NSVDist were more accurate than the models based on SeqA, Mondrian, or PAIS, especially for higher values of k.

We have tested the statistical significance of our results using the evaluation methodology recommended by [5]. First, we applied the non-parametric Friedman test to the null hypothesis that all anonymization algorithms (including the baseline majority rule classifier) provide the same classification accuracy across different values of k. Contrary to the standard Friedman test, which ranks different classification algorithms across different datasets, we have ranked different anonymization methods across different values of k. We have sorted the accuracy results in ascending order before ranking them so that the best method is the one with the highest average rank. Based on the p-values shown in Table 5 for each dataset, classification algorithm, and two different values of ℓ, the null hypothesis can be rejected at the level of 0.05 and higher for all examined cases, implying that the method of anonymization does have an impact on the classification accuracy of the model induced from generalized data. The average rankings of each anonymization method in every dataset are shown in Tables 7 and 8 for ℓ equal to one and ℓ greater than one, respectively.

Following the rejection of the null hypothesis by the Friedman test, we proceeded with the post-hoc Bonferroni-Dunn test to compare the NSVDist-based classifier to the best classifier from among the other four classifiers (the ones that correspond to the three remaining anonymization methods and the baseline majority classifier). This test was also repeated for each dataset, classification algorithm, and two different values of ℓ. Each p-value shown in Table 6 refers to the difference between NSVDist and the best of the remaining four anonymization methods (the one with the highest rank). Thus, it shows the largest value of p for each case. Obviously, when NSVDist outperforms the best of the other methods, it outperforms the remaining methods as well. The advantage of the NSVDist-based classifier over all other classifiers was found statistically significant (at the level of 0.05 and higher) in 31 cases out of 64. In additional 23 cases, it also provided the best performance, but the difference vs. the second best classifier was not significant statistically. Only in 10 cases, a different classifier significantly outperformed the NSVDist-classifier. However, from among those 10 cases, the winning classifier in 7 cases was the baseline majority classifier, and not one of the classifiers that were based on other anonymization algorithms. It is noteworthy that if we ignore the baseline majority classifier, the number of significant NSVDist 'wins' goes up to 41, whereas the number of its 'losses' goes down to 3 only. The detailed p-values of the post-hoc test comparing each one of the four alternative anonymization methods to NSVDist are shown in Tables 9 and 10 for ℓ equal to one and ℓ greater than one, respectively.



Dataset	div=1.0				div>1.0			
	W-J48	NBC	W-JRip	SVM	W-J48	NBC	W-JRip	SVM
Abalone	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Adult	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Breast Cancer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CMC	0.00	0.00	0.03	0.02	0.00	0.00	0.01	0.01
Ecoli	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00
Mammographic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Page Blocks	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
Yeast	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00

Table 5: Friedman Test: p-values

In our final set of experiments we compared the performance of NSVDist, in terms of classification accuracy, to that of the non-homogeneous anonymization algorithm (NHAA) of Wong et al. [43]. We selected the `Abalone` dataset since the global diversity of that dataset was the highest among all datasets (see Table 3), and that enabled us to conduct experiments with several values of the diversity ℓ. (Recall that unlike NSVDist, NHAA is limited to integer values of ℓ only). While NSVDist, as well as SeqA and Mondrian, can impose a conjunction of conditions — ℓ-diversity and k-anonymity with $k \geq \ell$, NHAA is guided solely by a diversity constraint; i.e, it can issue tables that are ℓ-diverse (and, consequently, are also k-anonymized with $k = \ell$). Hence, in the experiment set reported in Figure 5, where we compare NSVDist, NHAA and the Mondrian, we used in all experiments $k = \ell$.

Dataset	div=1.0				div>1.0			
	W-J48	NBC	W-JRip	SVM	W-J48	NBC	W-JRip	SVM
Abalone	**0.02**	0.19	**0.00**	**0.02**	**0.04**	0.15	**0.01**	**0.04**
Adult	**0.05**	**0.03**	0.07	**0.04**	**0.04**	**0.02**	0.07	**0.02**
Breast Cancer	**0.05**	**0.01**	0.07	**0.05**	0.88	0.67	0.28	0.44
CMC	**0.04**	**0.03**	**0.01**	**0.03**	**0.03**	**0.03**	0.08	**0.04**
Ecoli	**0.05**	**0.05**	**0.02**	0.88	0.16	0.09	0.20	0.99
Mammographic	**0.01**	**0.00**	**0.01**	**0.01**	0.19	**0.05**	0.28	0.12
Page Blocks	0.12	0.91	0.12	1.00	0.12	0.88	0.12	1.00
Yeast	0.12	**0.03**	0.31	0.88	0.12	0.06	0.37	0.88

Table 6: Bonferroni-Dunn test: p-values

Dataset	W-J48					NBC					W-JRip					SVM				
	NSVD.	PAIS	SeqA	Mond.	Majo.	NSVD.	PAIS	SeqA	Mond.	Majo.	NSVD.	PAIS	SeqA	Mond.	Majo.	NSVD.	PAIS	SeqA	Mond.	Majo.
Abalone	**5.00**	2.44	2.33	1.78	3.44	**4.67**	4.00	2.89	2.44	1.00	**5.00**	2.83	2.94	3.00	1.22	**5.00**	2.56	3.06	3.39	1.00
Adult	**5.00**	3.91	2.41	2.50	1.18	**5.00**	3.73	2.82	2.45	1.00	**5.00**	4.00	2.45	2.55	1.00	**5.00**	3.82	2.64	2.55	1.00
Breast Cancer	**5.00**	3.78	2.67	2.56	1.00	**5.00**	3.33	3.22	2.44	1.00	**5.00**	3.89	2.22	2.78	1.11	**4.78**	3.56	2.22	1.89	2.56
CMC	**4.63**	3.25	3.13	3.00	1.00	**4.75**	2.88	3.25	3.13	1.00	**4.63**	2.63	2.75	2.38	2.63	**4.63**	3.13	2.63	2.50	2.13
Ecoli	**4.86**	2.93	1.71	2.07	3.43	**4.86**	2.86	1.86	2.00	3.43	**5.00**	2.57	2.14	2.00	3.29	4.00	2.07	2.00	1.93	**5.00**
Mammographic	**4.89**	3.11	3.11	2.33	1.56	**5.00**	3.00	3.00	2.33	1.67	**4.89**	3.11	2.44	2.89	1.67	**5.00**	2.44	3.11	2.78	1.67
Page Blocks	**5.00**	4.00	1.71	1.43	2.86	3.86	3.14	1.71	1.29	**5.00**	**5.00**	4.00	1.43	1.57	3.00	2.29	3.14	2.14	2.71	**4.71**
Yeast	**5.00**	1.86	1.86	2.29	4.00	**5.00**	2.29	2.00	2.29	3.43	**4.71**	2.29	1.71	2.00	4.29	3.43	2.43	2.43	2.29	**4.43**

Table 7: Average ranking of each method ($div = 1.0$)

Dataset	W-J48					NBC					W-JRip					SVM				
	NSVD.	PAIS	SeqA	Mond.	Majo.	NSVD.	PAIS	SeqA	Mond.	Majo.	NSVD.	PAIS	SeqA	Mond.	Majo.	NSVD.	PAIS	SeqA	Mond.	Majo.
Abalone	**5.00**	2.44	2.44	1.44	3.67	**4.67**	3.89	3.11	2.33	1.00	**5.00**	3.00	3.11	2.56	1.33	**5.00**	2.22	3.67	3.11	1.00
Adult	**5.00**	3.82	2.64	2.55	1.00	**5.00**	3.64	2.55	2.82	1.00	**5.00**	4.00	2.73	2.27	1.00	**5.00**	3.64	2.77	2.59	1.00
Breast Cancer	3.67	**4.56**	2.89	2.89	1.00	4.06	**4.39**	3.00	2.56	1.00	**4.33**	3.89	3.22	2.44	1.11	**4.11**	4.00	1.78	1.44	3.67
CMC	**4.63**	3.13	3.13	3.13	1.00	**4.88**	3.38	3.00	2.75	1.00	**4.50**	3.13	1.88	2.13	3.38	**4.63**	2.88	2.50	1.75	3.25
Ecoli	**4.43**	3.00	1.50	2.50	3.57	**4.86**	2.86	1.71	1.86	3.71	**4.43**	2.86	1.86	2.14	3.71	3.00	2.57	2.14	2.29	**5.00**
Mammographic	**4.56**	3.89	2.00	1.67	2.89	**5.00**	3.78	1.89	1.56	2.78	**4.67**	4.22	1.67	1.78	2.67	**4.78**	3.89	1.56	1.44	3.33
Page Blocks	**5.00**	4.00	1.57	1.43	3.00	4.00	3.00	1.29	1.71	**5.00**	**5.00**	4.00	1.50	1.50	3.00	2.43	3.71	2.14	2.00	**4.71**
Yeast	**5.00**	2.14	1.57	2.29	4.00	**5.00**	2.57	2.14	1.57	3.71	**4.43**	2.14	2.00	2.29	4.14	3.71	2.43	2.00	2.14	**4.71**

Table 8: Average ranking of each method ($div > 1.0$)

The top plot in Figure 5 shows the LM information loss in the three algorithms for all values of $k = \ell > 1$ that the NHAA algorithm can accept for this dataset ($k = \ell = 2, 3, 4, 5$). NSVDist constantly yielded much lower information losses. The next four plots in Figure 5 show the accuracy of the corresponding classifiers. The NSVDist-trained classifier is always better than the other two. While the advantage of NSVDist over NHAA is not always significant, it should be noted that it does not suffer from the above mentioned limitations of the NHAA algorithm. (See a discussion of those limitations in Section 4.5).

5.4. Discussion

NSVDist produces anonymizations in which every single generalized record allows to link a given individual to a frequency distribution over the sensitive domain, where all

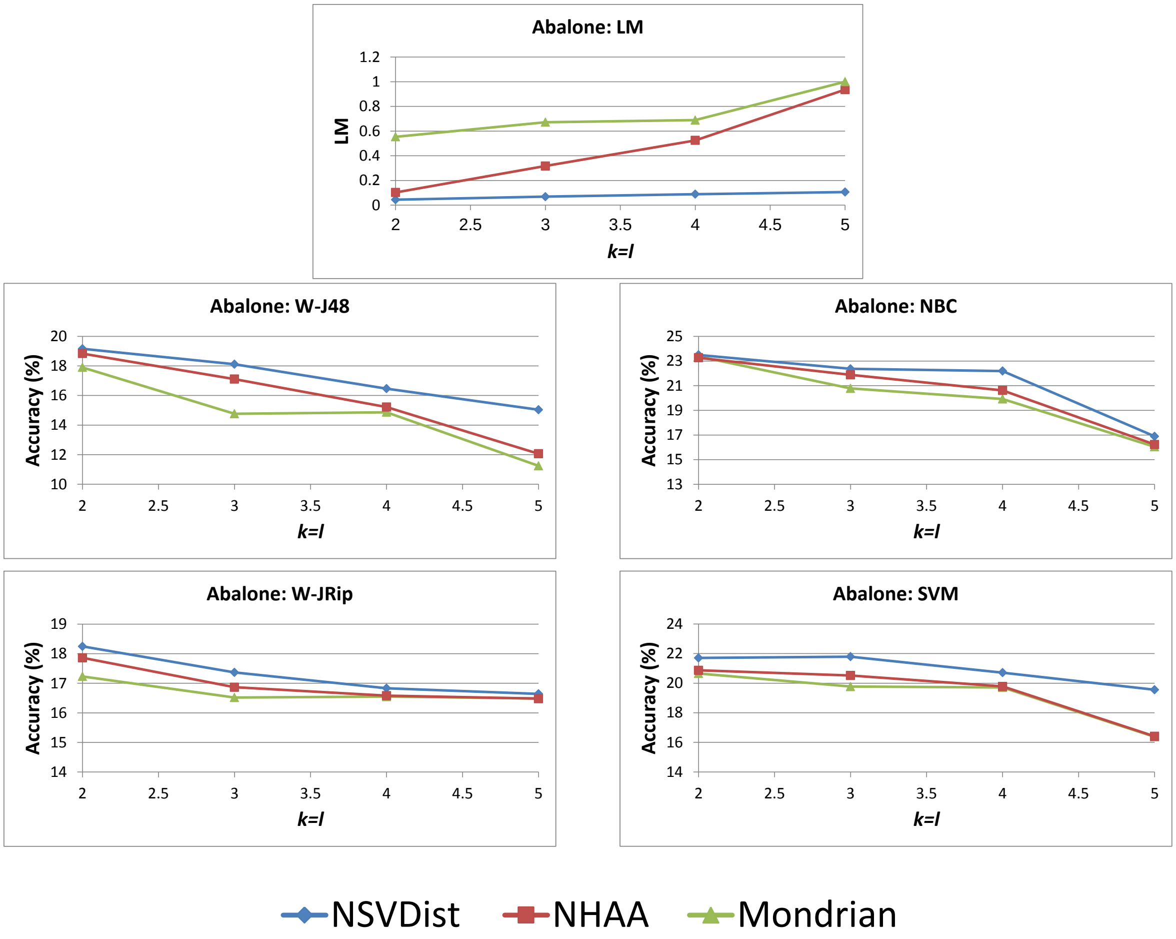


Figure 5: Comparison with the non-homogeneous anonymization algorithm (NHAA) of [43]

Dataset	W-J48				NBC				W-JRip				SVM			
	PAIS	SeqA	Mond.	Majo.	PAIS	SeqA	Mond.	Majo.	PAIS	SeqA	Mond.	Majo.	PAIS	SeqA	Mond.	Majo.
Abalone	0.00	0.00	0.00	0.02	0.19	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
Adult	0.05	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.04	0.00	0.00	0.00
Breast Cancer	0.05	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.07	0.00	0.00	0.00	0.05	0.00	0.00	0.00
CMC	0.04	0.03	0.02	0.00	0.01	0.03	0.02	0.00	0.01	0.01	0.00	0.01	0.03	0.01	0.00	0.00
Ecoli	0.01	0.00	0.00	0.05	0.01	0.00	0.00	0.05	0.00	0.00	0.00	0.02	0.01	0.01	0.01	0.88
Mammographic	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Page Blocks	0.12	0.00	0.00	0.01	0.20	0.01	0.00	0.91	0.12	0.00	0.00	0.01	0.84	0.43	0.69	1.00
Yeast	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.31	0.12	0.12	0.09	0.88

Table 9: Post-hoc test of each method ($div = 1.0$): p-values

Dataset	W-J48				NBC				W-JRip				SVM			
	PAIS	SeqA	Mond.	Majo.	PAIS	SeqA	Mond.	Majo.	PAIS	SeqA	Mond.	Majo.	PAIS	SeqA	Mond.	Majo.
Abalone	0.00	0.00	0.00	0.04	0.15	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.04	0.01	0.00
Adult	0.04	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.02	0.00	0.00	0.00
Breast Cancer	0.88	0.15	0.15	0.00	0.67	0.08	0.02	0.00	0.28	0.07	0.01	0.00	0.44	0.00	0.00	0.28
CMC	0.03	0.03	0.03	0.00	0.03	0.01	0.00	0.00	0.04	0.00	0.00	0.08	0.01	0.00	0.00	0.04
Ecoli	0.05	0.00	0.01	0.16	0.01	0.00	0.00	0.09	0.03	0.00	0.00	0.20	0.31	0.16	0.20	0.99
Mammographic	0.19	0.00	0.00	0.01	0.05	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.12	0.00	0.00	0.03
Page Blocks	0.12	0.00	0.00	0.01	0.12	0.00	0.00	0.88	0.12	0.00	0.00	0.01	0.94	0.37	0.31	1.00
Yeast	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.06	0.00	0.00	0.01	0.37	0.06	0.02	0.03	0.88

Table 10: Post-hoc test of each method ($div > 1.0$): p-values

sensitive values have frequency no larger than $1/\ell$. Hence, any single generalized record enables an adversary to learn the sensitive value of his target individual with probability at most $1/\ell$. However, a large collection of generalized records reveals more information than a single generalized record, since the adversary can use that collection to learn a classifier and then deduce his target individual's sensitive data with probability higher than the intended $1/\ell$. For example, Figure 3 shows that the anonymization of `Mammographic` records using $\ell = 1.31$ — an anonymization that aims at upper bounding the sensitive inference probability by $1/\ell = 76.3\%$ — enables to infer a classifier whose sensitive inference success probability may be as high as 80%.

A closely related finding was reported by Kifer in [20]. He showed that it is possible to extract from ℓ-diverse tables belief probabilities greater than $1/\ell$ by means of the so-called deFinetti attack. That attack uses the anonymized table in order to learn a classifier that, given the quasi-identifier record of an individual in the underlying population, is able to predict the corresponding sensitive value with probability greater than the intended $1/\ell$ bound.

The question that arises from both our study and Kifer's is as follows: assume that an adversary uses the anonymized data in order to learn a classifier and, subsequently, achieves belief probabilities regarding the sensitive data of individuals that are higher than the intended bounds. Does that constitute a privacy breach? Stated differently, the question is whether the inference of private sensitive data about specific individuals from the general behavior of the population, can be regarded as a privacy breach.

To answer this question positively, the success of the attack when launched against records that are part of the original table T (which was generalized, published and then used to learn a classification model) must be shown to be significantly higher than its success against records that are not part of that table. Figure 6 shows the accuracy of the Naïve Bayes classifier that was induced from an anonymization of the `Mammographic` dataset with $\ell = 1.5$ and various values of k. It shows the accuracy of that classifier over the original (non-generalized) training data records (i.e., the records of T) that were given to NSVDist as input, compared with its accuracy on the non-generalized testing data records. As can be seen clearly, the two curves almost coincide. We have obtained very similar results for the other anonymized datasets and the other classifiers as well. Namely, even though we published data on Alice, Bob, and Carol, the above described "classifier attack" presents a similar level of risk for them, as well as for David, Elaine, and Frank who were not included in the original table T that was used to generate the published anonymized data and, subsequently, to learn the classification model. Hence, such an "attack" cannot be regarded as a breach of privacy. It can only be regarded as a successful learning of the behavior of the general population, which is the raison d'être of any data publishing.

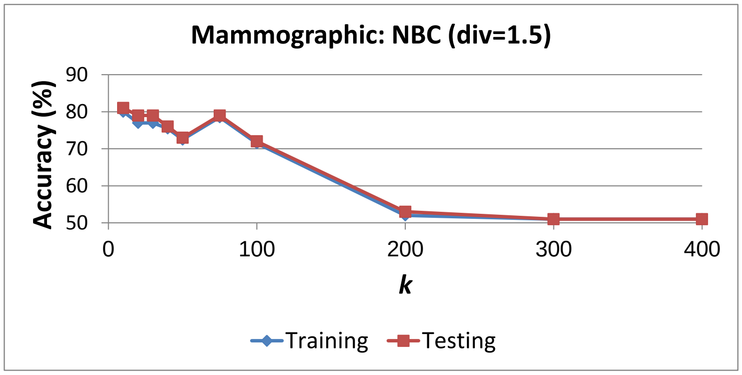


Figure 6: Classification performance: Original training and testing data

6. Conclusions

In this paper we presented a new privacy-preserving data publishing algorithm called NSVDist (Non-homogeneous generalization with Sensitive Value Distributions). That algorithm is based on non-homogeneous anonymization of the quasi-identifiers, coupled with the generalization of the sensitive values into frequency distributions. Since that algorithm is characterized by smaller information losses than leading anonymization algorithms, our research hypothesis was that the proposed algorithm allows the data owner to release the data in a more secure form (represented by a higher value of k) while expecting the data miner to induce accurate classification models from the published data. Our experimental results confirm that hypothesis in most cases. Those findings suggest that the framework of non-homogeneous anonymizations, which allows lower information losses, might be more adequate for data mining purposes than homogeneous anonymizations.

Directions for future research include the following:

(a) Experimentation with additional data mining algorithms such as clustering or association rules.

(b) In this paper, we studied the simplest case of a single sensitive attribute, which is also a classification attribute. The proposed approach to non-homogeneous anonymization can be extended to more general cases like disjoint or partially overlapping sets of several sensitive and classification attributes.

(c) Extending the NSVDist algorithm for the case of a sequential release of data attributes [29, 37, 41]. In the sequential release scenario, several releases of the same table are published over a period of time, where each release contains a different set of the table attributes, as dictated by the purposes of the release. The goal is to protect the private information from adversaries who examine the entire sequential release.

Acknowledgments.

We thank Dr. Benjamin C. M. Fung for providing us with the PAIS software.

[1] Agrawal, R. and Srikant, R. 2000. Privacy-preserving data mining. In *The ACM SIGMOD International Conference on Data Management (SIGMOD)*. 439–450.

[2] Bayardo, R. and Agrawal, R. 2005. Data privacy through optimal k-anonymization. In *International Conference on Data Engineering (ICDE)*. 217–228.

[3] Burnett, L., Barlow-Stewart, K., Proos, A., and Aizenberg, H. 2003. The" GeneTrustee": a universal identification system that ensures privacy and confidentiality for human genetic databases. *Journal of Law and Medicine 10,* 4, 506.

[4] Cohen, W. W. 1995. Fast effective rule induction. In *International Conference on Machine Learning (ICML)*. 115–123.

[5] Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research 7*, 1–30.

[6] Frank, A. and Asuncion, A. 2010. UCI machine learning repository.

[7] Fung, B., Wang, K., Chen, R., and Yu, P. 2010. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys (CSUR) 42*, 1–53.

[8] Fung, B. C. M., Wang, K., and Yu, P. S. 2005. Top-down specialization for information and privacy preservation. In *International Conference on Data Engineering (ICDE)*. 205–216.

[9] Fung, B. C. M., Wang, K., and Yu, P. S. 2007. Anonymizing classification data for privacy preservation. *IEEE Transactions on Knowledge and Data Engineering 19,* 5, 711–725.

[10] Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. 2009. A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Transactions on Database Systems 34*, 1–47.

[11] Gionis, A., Mazza, A., and Tassa, T. 2008. k-Anonymization revisited. In *International Conference on Data Engineering (ICDE)*. 744–753.

[12] Gionis, A. and Tassa, T. 2009. k-Anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering 21*, 206–219.

[13] Goldberger, J. and Tassa, T. 2010. Efficient anonymizations with enhanced utility. *Transactions on Data Privacy 3*, 149–175.

[14] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter 11,* 1, 10–18.

[15] Han, J., Kamber, M., and Pei, J. 2011. *Data Mining: Concepts and Techniques*, third ed. Morgan Kaufmann.

[16] Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second ed. Springer Verlag.

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

[18] Iyengar, V. 2002. Transforming data to satisfy privacy constraints. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).* 279–288.

[19] Kenig, B. and Tassa, T. 2012. A practical approximation algorithm for optimal k-anonymity. *Data Mining and Knowledge Discovery 25*, 134–168.

[20] Kifer, D. 2009. Attacks on privacy and definetti's theorem. In *The ACM SIGMOD International Conference on Data Management (SIGMOD).* 127–138.

[21] Kisilevich, S., Rokach, L., Elovici, Y., and Shapira, B. 2010. Efficient multidimensional suppression for k-anonymity. *IEEE Transactions on Knowledge and Data Engineering 22,* 3, 334–347.

[22] Kryszkiewicz, M. 1998. Rough set approach to incomplete information systems. *Information Sciences 112,* 1–4, 39–49.

[23] Kryszkiewicz, M. 1999. Rules in incomplete information systems. *Information Sciences 113,* 3–4, 271–292.

[24] LeFevre, K., DeWitt, D., and Ramakrishnan, R. 2005. Incognito: efficient full-domain k-anonymity. In *The ACM SIGMOD International Conference on Data Management (SIGMOD).* 49–60.

[25] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. 2006a. Mondrian multidimensional k-anonymity. In *International Conference on Data Engineering (ICDE).* 25.

[26] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. 2006b. Workload-aware anonymization. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).* 277–286.

[27] Li, N., Li, T., and Venkatasubramanian, S. 2007. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *International Conference on Data Engineering (ICDE).* 106–115.

[28] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. 2006. l-Diversity: privacy beyond k-anonymity. In *International Conference on Data Engineering (ICDE)*. 24.

[29] Matatov, N., Rokach, L., and Maimon, O. 2010. Privacy-preserving data mining: a feature set partitioning approach. *Information Sciences 180,* 14, 2696–2720.

[30] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. 2006. Yale: rapid prototyping for complex data mining tasks. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 935–940.

[31] Mohammed, N., Chen, R., Fung, B. C. M., and Yu, P. S. 2011. Differentially private data release for data mining. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 493–501.

[32] Mohammed, N., Fung, B. C. M., Hung, P. C. K., and kwong Lee, C. 2009. Anonymizing healthcare data: a case study on the blood transfusion service. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 1285–1294.

[33] Nergiz, M. E. and Clifton, C. 2007. Thoughts on k-anonymization. *Data & Knowledge Engineering 63*, 622–645.

[34] Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

[35] Samarati, P. 2001. Protecting respondent's privacy in microdata release. *IEEE Transactions on Knowledge and Data Engineering 13*, 1010–1027.

[36] Samarati, P. and Sweeney, L. 1998. Generalizing data to provide anonymity when disclosing information. In *The ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*. 188.

[37] Shmueli, E., Tassa, T., Wasserstein, R., Shapira, B., and Rokach, L. 2012. Limiting disclosure of sensitive data in sequential releases of databases. *Information Sciences 191*, 98–127.

[38] Sweeney, L. 2002. k-Anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10,* 5, 557–570.

[39] Tassa, T., Mazza, A., and Gionis, A. 2012. k-Concealment: An alternative model of k-type anonymity. *Transactions on Data Privacy 5*, 189–222.

[40] Truta, T., Campan, A., and Meyer, P. 2007. Generating microdata with p-sensitive k-anonymity property. In *Secure Data Management*. 124–141.

[41] Wang, K. and Fung, B. 2006. Anonymizing sequential release. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 414–423.

[42] Wong, R., Li, J., Fu, A., and Wang, K. 2006. (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 754–759.

[43] Wong, W. K., Mamoulis, N., and Cheung, D. W.-L. 2010. Non-homogeneous generalization in privacy preserving data publishing. In *The ACM SIGMOD International Conference on Data Management (SIGMOD)*. 747–758.

[44] Xiao, X. and Tao, Y. 2006. Anatomy: Simple and Effective Privacy Preservation. In *The International Conference on Very Large Databases (VLDB)*. 139–150.

Appendix A. Approximation guarantee of Algorithm 1.

Algorithm 1 replaces each record $R_n \in T$ with a generalized record $\overline{R}_n$ which is the closure of R_n and additional $k-1$ records from T. The $k-1$ additional records are selected so that the resulting information loss is as small as possible and the diversity requirement is respected. One possible approach towards carrying out that computation is to scan, for each record $R_n \in T$, all $\binom{N-1}{k-1}$ possible selections of $k-1$ masking records and then choose those $k-1$ records that yield a closure with sufficient diversity and minimal information loss. However, that approach is impractical since its overall runtime is $O(N^k)$. The approach implemented in Algorithm 1 offers a practical alternative. We show below that under two reasonable assumptions on the information loss measure, Algorithm 1 outputs an anonymization in which the information loss approximates the minimal possible one to within a multiplicative factor of $k-1$. We proceed to state and prove this theoretical result.

Definition. *A measure of information loss $IL(\cdot)$ is monotone if $IL(\overline{R}) \le IL(\overline{R}')$ whenever $\overline{R} \sqsubseteq \overline{R}'$ (namely, if further generalizations cannot decrease the information loss). The measure $IL(\cdot)$ is sub-additive if for any two subsets of records $B_1, B_2 \subset A_1 \times \cdots \times A_{M+1}$ that have a nonempty intersection, $IL(B_1 \cup B_2) \le IL(B_1) + IL(B_2)$.*

Theorem. *Let $\overline{T} = \{\overline{R}_1, \ldots, \overline{R}_N\}$ be the (k,ℓ)-anonymization of T as issued by Algorithm 1 and let $\overline{T}' = \{\overline{R}'_1, \ldots, \overline{R}'_N\}$ be a non-homogeneous (k,ℓ)-anonymization with sensitive value distributions in which the information loss is minimal. Assume that the information loss measure IL is monotone and sub-additive. Then $IL(\overline{R}_n) \le (k-1) \cdot IL(\overline{R}'_n)$ for all $n \in [N]$.*

Proof. For a fixed $n \in [N]$, let $R_{i_1}, \ldots, R_{i_{k-1}}$ be the records that were selected by Algorithm 1 in order to compute $\overline{R}_n$ as the closure of $\{R_n, R_{i_1}, \ldots, R_{i_{k-1}}\}$. The indices of those records denote the order in which they were selected; i.e., R_{i_j} is the record that was selected in the jth application of Step 5 of the algorithm for the seed record R_n. $\overline{T}' = \{\overline{R}'_1, \ldots, \overline{R}'_N\}$ is a (k,ℓ)-anonymization of T for which the information loss is minimal. By the k-anonymity property, $\overline{R}'_n$ is consistent with R_n and $k-1$ other records. We claim, and prove later, that there exists an ordering of those $k-1$ records, say $R_{i'_1}, \ldots, R_{i'_{k-1}}$, for which the following inequality holds,

$$IL(\{R_n, R_{i_1}, \ldots, R_{i_{j-1}}, R_{i_j}\}) \le \\ IL(\{R_n, R_{i_1}, \ldots, R_{i_{j-1}}, R_{i'_j}\})\,. \tag{A.1}$$

Since, by sub-additivity,

$$IL(\{R_n, R_{i_1}, \ldots, R_{i_{j-1}}, R_{i'_j}\}) \le \\ IL(\{R_n, R_{i_1}, \ldots, R_{i_{j-1}}\}) + IL(\{R_n, R_{i'_j}\})\,, \tag{A.2}$$

we conclude, by (A.1) and (A.2), that

$$\begin{aligned} & IL(\{R_n, R_{i_1}, \ldots, R_{i_{j-1}}, R_{i_j}\}) \leq \\ & \qquad\qquad IL(\{R_n, R_{i_1}, \ldots, R_{i_{j-1}}\}) + IL(\{R_n, R_{i'_j}\})\,. \end{aligned} \tag{A.3}$$

Applying inequality (A.3) repeatedly for $j = k-1$ down to $j = 1$ we infer that

$$IL(\{R_n, R_{i_1}, \ldots, R_{i_{k-1}}\}) \leq \sum_{j=1}^{k-1} IL(\{R_n, R_{i'_j}\})\,. \tag{A.4}$$

Monotonicity implies that for all $1 \leq j \leq k-1$

$$IL(\{R_n, R_{i'_j}\}) \leq IL(\{R_n, R_{i'_1}, \ldots, R_{i'_{k-1}}\})\,. \tag{A.5}$$

Hence, by (A.4) and (A.5),

$$\begin{aligned} & IL(\{R_n, R_{i_1}, \ldots, R_{i_{k-1}}\}) \leq \\ & \qquad\qquad (k-1) \cdot IL(\{R_n, R_{i'_1}, \ldots, R_{i'_{k-1}}\})\,. \end{aligned} \tag{A.6}$$

Since the left hand side in (A.6) equals $IL(\overline{R}_n)$ while the right hand side equals $(k-1) \cdot IL(\overline{R}_n^{'})$, we conclude that $IL(\overline{R}_n) \leq (k-1) \cdot IL(\overline{R}_n^{'})$.

We now turn to prove inequality (A.1). Let $A = \{R_{i_1}, \ldots, R_{i_{k-1}}\}$ be the set of $k-1$ records that were selected in Step 5 of the algorithm in order to generate the generalization $\overline{R}_n$ of R_n. Let A' be the set of $k-1$ records with which $\overline{R}_n^{'}$ (the generalization of R_n in the (k, ℓ)-anonymization $\overline{T}^{'}$ with minimal information loss) is consistent. We proceed to induce an ordering of A' in the following manner:

(I) Records in A' that appear also in A will get the same index as they have in A. For example, if a record in A' equals the third record in A, i.e. $\overline{R}_{i_3}$, we shall denote it $\overline{R}_{i'_3}$. Records of that type will be called records of type I.

(II) Let B and B' be the subsets of records of A and A' that are not of type I. If there exists in B' a record with a sensitive value that appears also in a record in B, we assign the former the same index as the latter. For example, if a record in B' has the same sensitive value as $\overline{R}_{i_4} \in B$, we shall denote it by $\overline{R}_{i'_4}$. Records of that type will be called records of type II.

(III) Assume that all records of type II were identified and indexed. All remaining records in A' will be called records of type III. We assign them the remaining indices from $i'_1, \ldots, i'_{k-1}$ that were not assigned so far, in an arbitrary manner.

To illustrate the process, assume that $k = 6$ and that for R_1 we have

$$A = \{(R_7, a), (R_{20}, b), (R_3, a), (R_9, c), (R_5, b)\} .$$

Namely, when applying Algorithm 1 on R_1, the selected records were R_7 (whose sensitive value is a), then R_{20} (with sensitive value b) and so forth. Assume further that

$$A' = \{(R_3, a), (R_7, a), (R_8, c), (R_{15}, c), (R_{24}, d)\} .$$

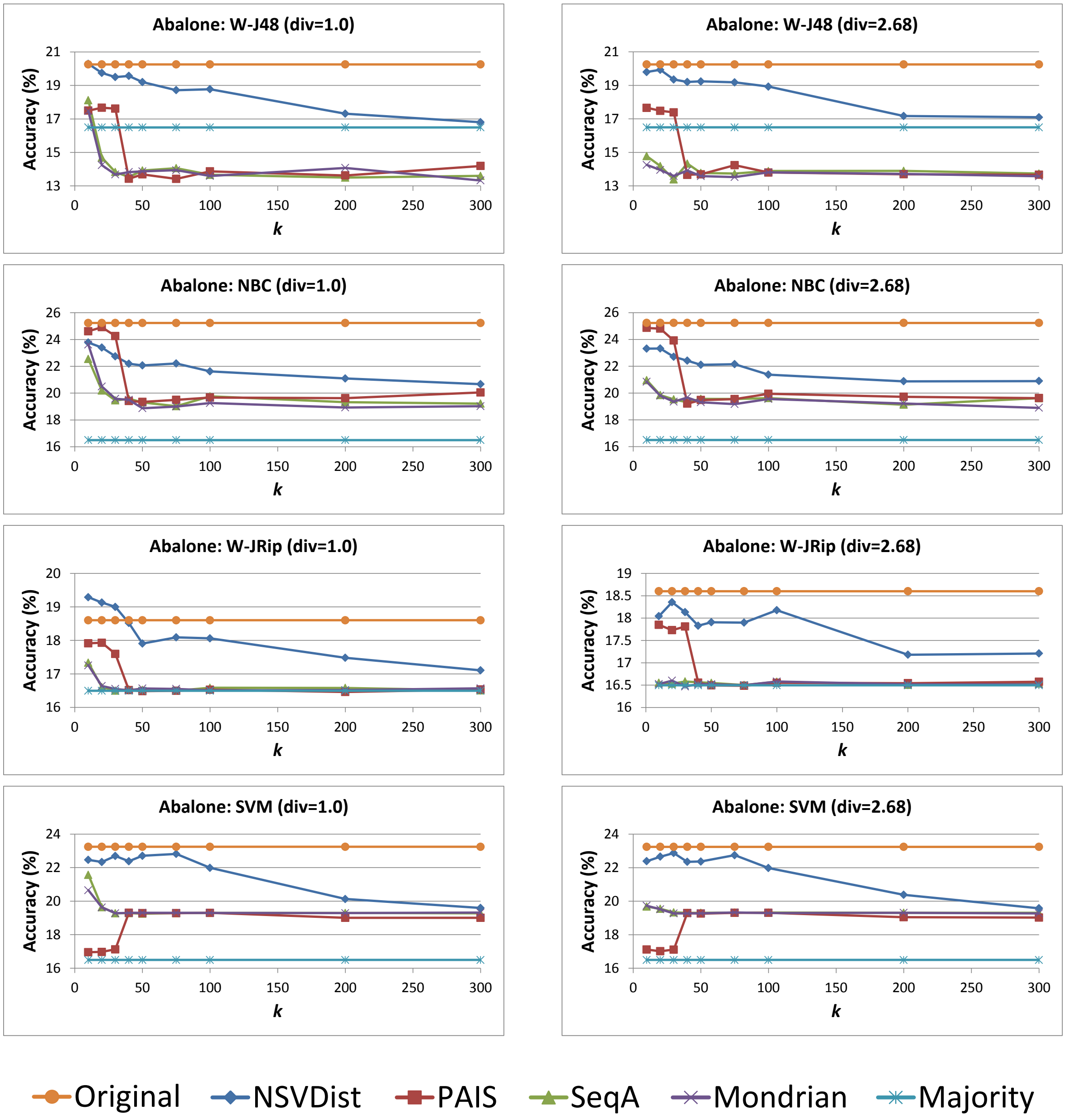
The ordering of the records in A' is done as follows: First, we place R_7 in the first place and R_3 in the third place, in accord with their position in A. Out of the remaining records, R_8 has a sensitive value that appears in the remaining records in A; so we place it in the fourth position (since the fourth record in A has the same sensitive value of c). We are left with R_{15} and R_{24}, which we place in the remaining positions — second and fifth. The resulting order is

$$A' = \{(R_7, a), (R_{15}, d), (R_3, a), (R_8, c), (R_{24}, e)\} .$$

Here, R_3 and R_7 are records of type I, R_8 is of type II, and R_{15} and R_{24} are of type III.

We claim that with this ordering, $R_{i'_j}$ was a legitimate candidate in the stage where R_{i_j} was selected, $1 \leq j \leq k - 1$. Since the algorithm selected R_{i_j} over $R_{i'_j}$, we infer that inequality (A.1) must hold. Indeed, if $R_{i'_j}$ is of type I, inequality (A.1) holds in a trivial manner. If $R_{i'_j}$ is of type II, then it must have been a legitimate record to select, since the record R_{i_j} that was selected eventually has the same sensitive value. Also if $R_{i'_j}$ is of type III then it was a legitimate candidate at the stage when R_{i_j} was selected (since any record with a sensitive value that still does not appear in an ℓ-diverse set can be added to that set without violating the ℓ-diversity condition). Hence, inequality (A.1) holds and the proof is thus complete. □

Appendix B. Experiments with Additional Datasets



Abalone: W-J48 (div=1.0)

Abalone: W-J48 (div=2.68)

Abalone: NBC (div=1.0)

Abalone: NBC (div=2.68)

Abalone: W-JRip (div=1.0)

Abalone: W-JRip (div=2.68)

Abalone: SVM (div=1.0)

Abalone: SVM (div=2.68)

Figure B.7: Classification performance using anonymized data (Abalone)

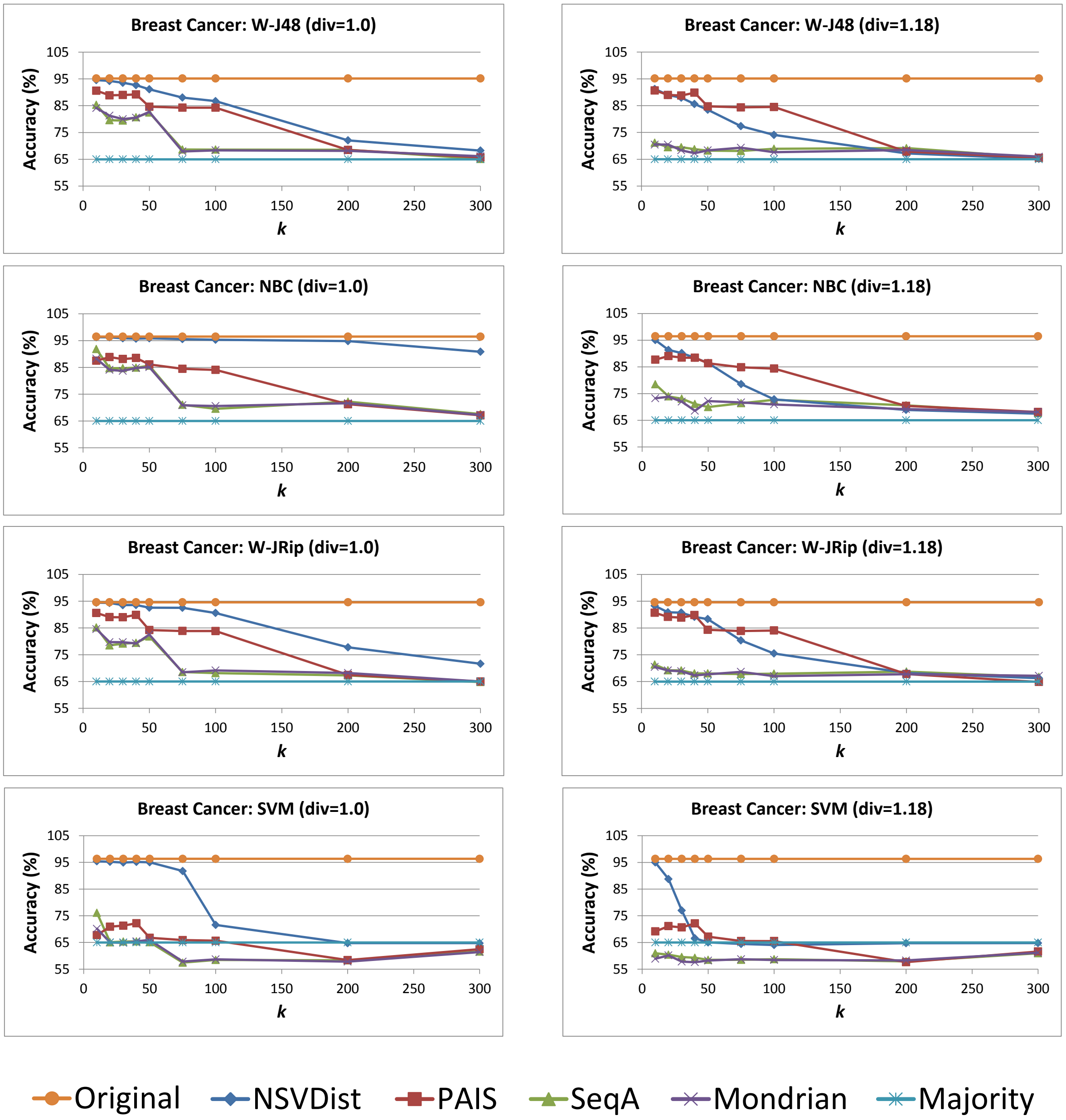


Figure B.8: Classification performance using anonymized data (Breast Cancer)

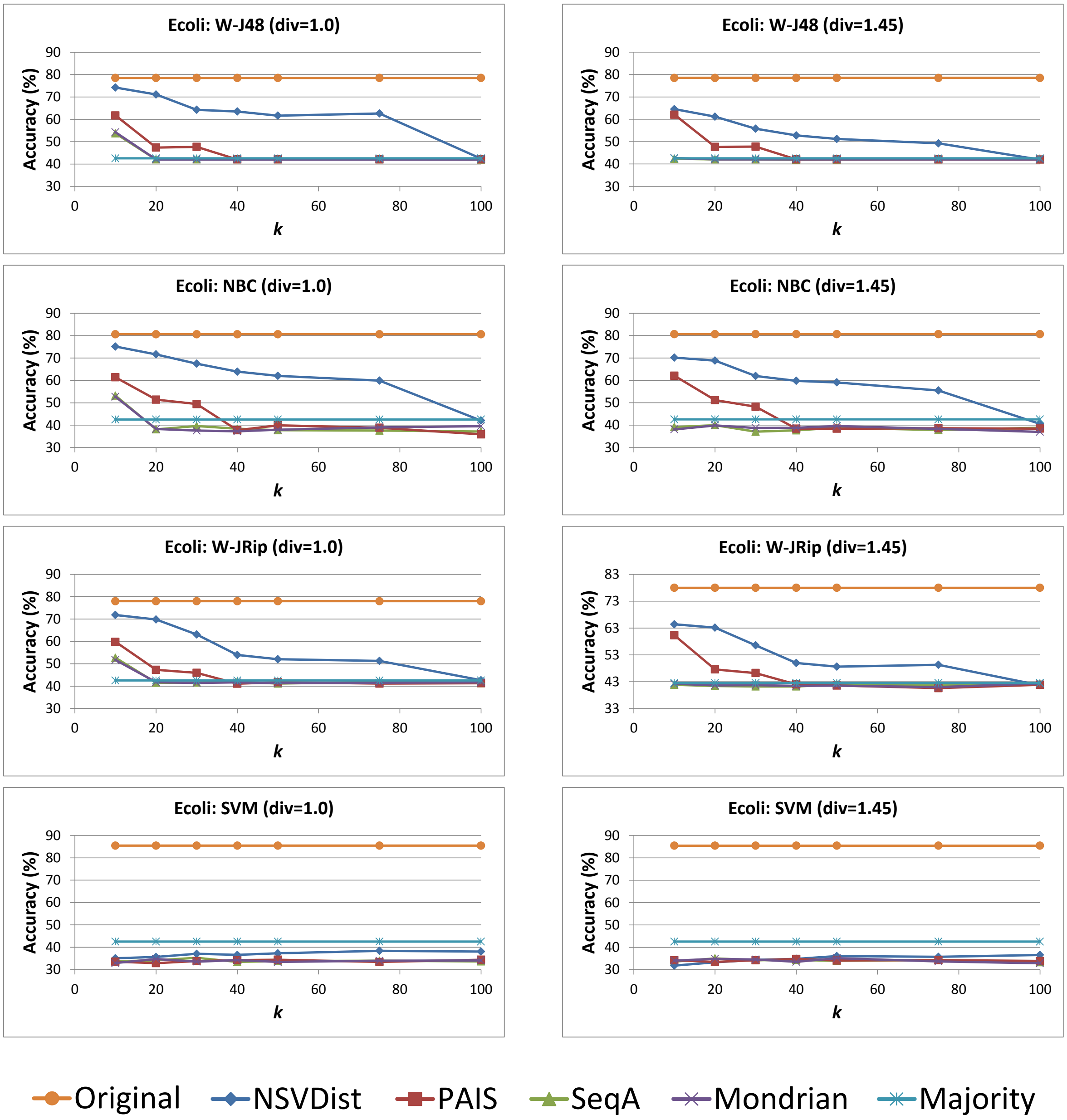


Figure B.9: Classification performance using anonymized data (Ecoli)

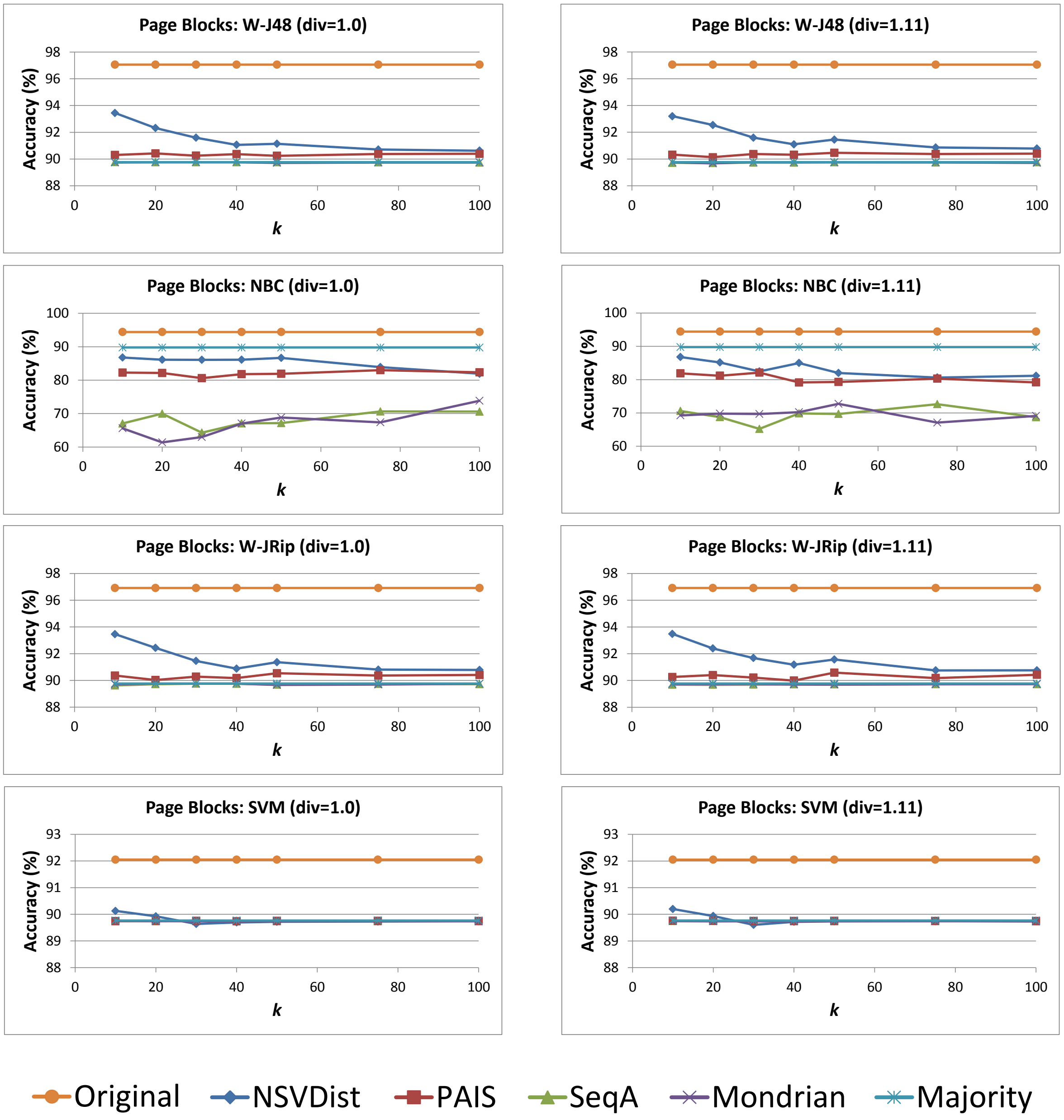


Figure B.10: Classification performance using anonymized data (Page Blocks)

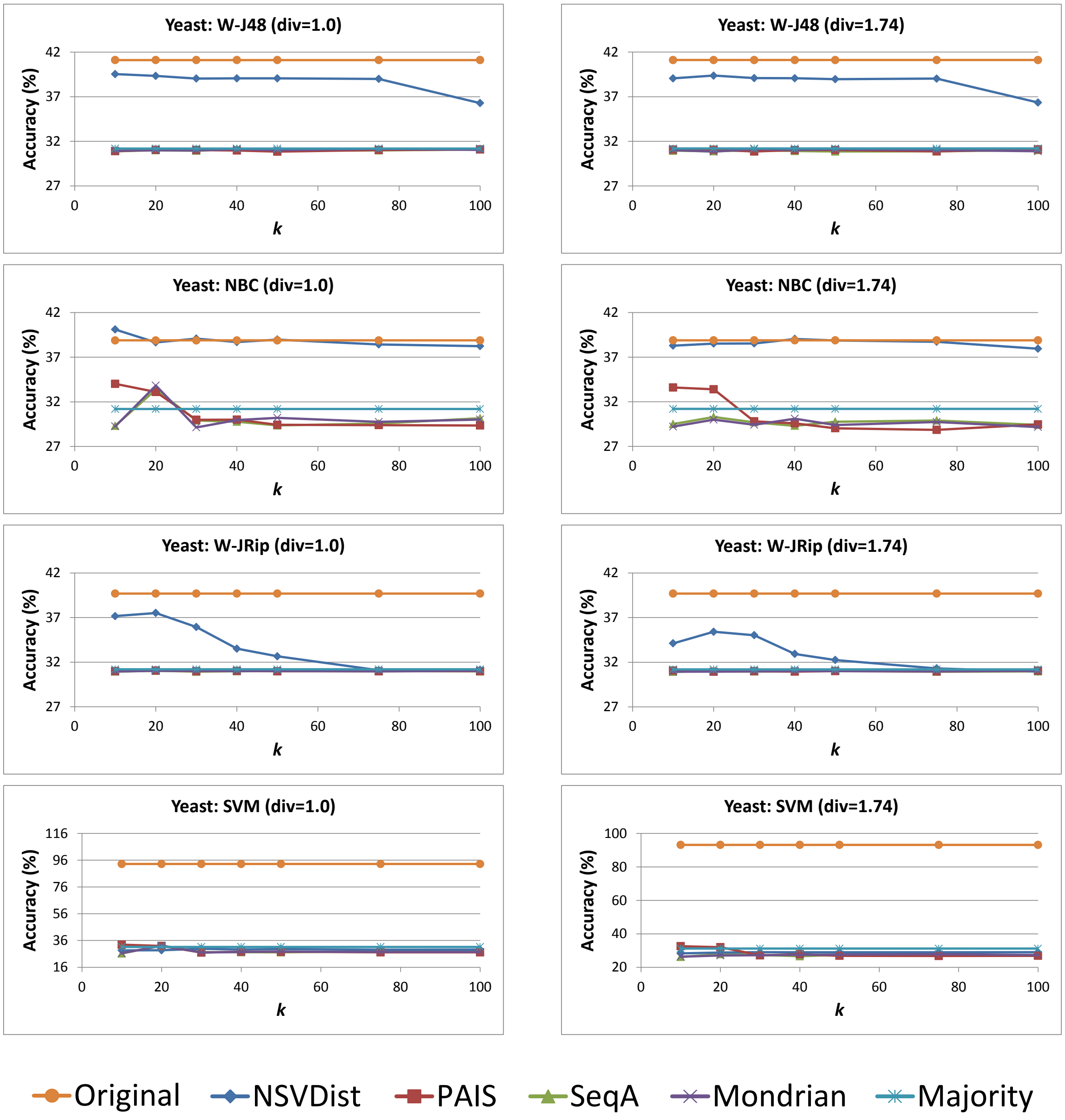


Figure B.11: Classification performance using anonymized data (Yeast)