PSD 2020, LNCS 12276
This is a self-archived pre-print version of this article.
The final publication is available at Springer via
https://doi.org/10.1007/978-3-030-57521-2_9.

An Analysis of Different Notions of Effectiveness in k-Anonymity

Tanja Šarčević[1[0000−0003−0896−9193]], David Molnar[2], and Rudolf Mayer[1,2[0000−0003−0424−5999]]

[1] SBA Research, Vienna, Austria
[2] Vienna University of Technology, Vienna, Austria
{TSarcevic,RMayer}@sba-research.org

Abstract. k-anonymity is an approach for enabling privacy-preserving data publishing of personal, sensitive data. As a result of this anonymisation process, the utility of the sanitised data is generally lower than on the original data. Quantifying this utility loss is therefore important to estimate the usefulness of the resulting datasets. In this paper, we analyse several of these utility aspects.
Data utility can be measured as a direct property of the resulting, anonymised dataset, or via the effectiveness that a statistical analysis, such as a machine learning model, achieves upon this dataset, as compared to the original data. While the latter is more tailored to the specific dataset, it is also generally less efficient. We therefore analyse whether there is a correlation between these two types of measures, and whether the measurement on the effectiveness can be substituted by a measurement of the data properties. Further, we evaluate to what extent different solutions for the same level of k-anonymity differ in regards to effectiveness.

Keywords: k-anonymity · Utility evaluation · Utility Metrics · Machine Learning

1 Introduction

Day after day we generate more and more data in every sector of our daily life. There are different types of data regarding the domain, but one of the most valuable is personal data, since they contain information about people, which is relevant for commercial as well as other purposes, such as healthcare.

For any statistical analysis such data is the key ingredient. However, individuals' privacy can be compromised even if direct personal identifiable information is removed. The Netflix Prize from 2007 is a famous example of how customer privacy can be threatened without any identifiers, by matching two related datasets.

Distributing personal data is highly regulated by law, especially so in the European Union with the General Data Protection Regulation (GDPR), which came into effect in May 2018. For many purposes, datasets have to be therefore anonymised before distributing them, in order to avoid identification of the people contained in these datasets.

k-Anonymity is a privacy model that can be applied to sensitive datasets by obfuscating information that can be utilised to re-identify individual records in a dataset from which direct identifiers have been removed. Besides weaknesses in the privacy guarantee of k-anonymity, and other proposed models such as Differential Privacy [1], or also approaches such as synthetic data generation [2], these also do have practical downsides. k-anonymity as a model that facilitates easy data sharing is thus still considered in several settings.

Another aspect to consider for datasets that have been such treated is the utility of the resulting data. While anonymisation techniques such as k-anonymity, as well as its extensions such as l-diversity and t-closeness, provide individuals' anonymity in datasets, it at the same time disturbs the effectiveness of machine learning algorithms. This is due to that, when sanitising a dataset, via anonymisation or other approaches, some sensitive information at the level of individual records is invariably removed [3].

Given a candidate for anonymised data, a *utility metric* quantifies the utility (or sometimes called the quality) of this release candidate (resp. the information loss due to the anonymisation process). Data utility can in principal be evaluated via two approaches. One is to utilise one or more quantitative measures of information loss (see [3] and 2 for an overview). Another approach is to measure the effectiveness of the final statistical analysis to be carried out on the data, such as a predictive machine learning model, compared to an analysis that would have been using the original, unabridged data. The latter is a very task-specific approach, and further less efficient, as it is generally more resource consuming (time, computing power, etc.) than the quantitative measures on the data itself.

We are therefore specifically interested in to what extent these two approaches correlate, and whether one can be used as a proxy measure for the other. We are estimating this in an experimental evaluation, utilising different machine learning models on different classification tasks. We thus utilise correlation analysis to find relationships between classifier behaviour and utility metrics of the anonymised datasets. As part of this evaluation, we generally compare the utility of the anonymised datasets to the original, source data.

Another aspect of our investigation is centred along the fact that there is generally not only one solution for achieving a certain sanitised version of an original dataset that fulfils the desired level of k-anonymity. In contrast, often a large number of candidate solutions exists, and finding the optimal solution is generally solved via heuristic approaches. Therefore, most algorithms utilise implicitly some data utility metric when deciding which solution to find. We want to investigate to what extent this influences the utility of the final, resulting dataset. To this end, we carry out experiments not only on the "best" found candidate, but also on different candidates covering the entire range of the solution space.

The remainder of this paper is organised as follows. Section 2, before Section 3 will detail our evaluation methodology. Section 4 discusses and analyses our results, before Section 5 provides conclusions and future work.

2 Related Work

The concept of k-anonymity was first introduced in the paper of Samarati and Sweeney [4]. This privacy model can be used to obfuscate sensitive datasets in order to be able to share them with other parties, thus also fulfilling regulations such as the EU's GDPR by anonymising data.

In a dataset, we can generally distinguish different types of attributes (sometimes called variables, or features). On the one hand, *(directly) identifying attributes* directly reveal the identity of a data record. Examples are the full name (to some extent), an e-mail address, or a social security number. As a general pre-processing steps, these are in practices removed from the dataset before publishing, or at least replaced with a pseudonym as identifier.

Quasi-identifiers (QIs) do not directly identify a person, but may become uniquely identifying when used in combination with other quasi-identifiers. An example can be a date of birth in combination with information on the residence of a person, even if in the relatively coarse form of a ZIP code. It has to be noted that this will not apply for all records in the dataset, but in some settings, a large number of them can become re-identifiable. For instance, [5] mentions that 87% of U.S. citizens in 2002 could be re-identified by using attributes zip code, sex and date of birth.

Besides potentially helping in identification, quasi-identifiers often hold significant, demographic information, which is required in analysis processes for differentiating between different groups. In medical analysis, for example, it is often important to differentiate between age groups, the type of job, or information on the location of the residence of patients. Thus, this information cannot simply be omitted as well.

Sensitive data is contained in attributes that for example hold information about a certain type of illness, or the salary of an individual. These are generally the main target in statistical analysis, and can therefore not be omitted or obfuscated.

k-anonymity is a property of the dataset, which ensures that for the identified quasi-identifiers, there are at least k records in the dataset that are indistinguishable in regards to the quasi identifiers. These records that share the same quasi-identifier values are called equivalence groups (or classes) or Q-blocks.

k-anonymity can be achieved by suppression and generalisation, where by suppression we mean simple deletion of values, whereas generalisation refers to a decrease in a value's granularity.

Generalisation utilises so-called *generalisation hierarchies*, which run from leaf nodes denoting particular values via internal nodes to their most general root. In the generalisation process for k-anonymity, one traverses the tree from a leaf node of the original input value upwards until we can construct an equivalence group with all quasi-identifiers being duplicates of one another.

One further needs to distinguish between a global or local generalisation. Global generalisation means that an attribute is put to the same generalisation level for each data record. Local generalisation on the other hand optimises the generalisation by choosing a minimal required loss of precision for

each equivalence group. As each level of generalisation invokes an increasing loss of specificity, we want to minimise a dataset's overall information loss. This makes k-anonymisation an NP-hard problem due to an exponential number of possible data-row combinations one can examine. For local generalisation, the search space becomes even larger.

Based on k-anonymity, several related concepts have been proposed, each addressing potential attack vectors for disclosure that the original model did not consider. l-diversity [6] and t-closeness [7] are among the most prominent of those, ensuring diversity among the sensitive attributes. We do not evaluate these at this stage of our work, however.

Several different, mostly heuristic, approaches have been proposed for finding an optimal level of suppression and generalisation for achieving a specific level of k-anonymity. Samarati [8] introduces a concept of *minimal generalisation* that captures the property of the release process not to distort the data more than needed to achieve k-anonymity. One globally-optimal anonymisation algorithm is Flash [9], which we utilise in the implementation provided with the anonymisation software ARX [3]. We further utilise an algorithm providing local generalisation, using a version of a greedy clustering algorithm called SaNGreeA (Social network greedy clustering), [10], as implemented for relational data by [11].

Measuring the quality of the output datasets is a complex aspect. It can be addressed by supporting multiple quality models which can be used as objective function in the optimisation process of the output data. These can include cell-oriented, attribute-oriented and record-oriented general-purpose models.

In the Flash algorithm we utilise, the default objective function of the anonymisation process is *Loss*, which " summarises the degree to which transformed attribute values cover the original domain of an attribute." [4] Since the anonymisation is based on this metric, the prime interest is the correlation of this measurement with the classification results. However, we further compute additional utility metrics that describe the output dataset, namely:

- *Record-level squared error:* This utility metric is the sum of squared errors in groups of indistinguishable records in the transformed dataset. The error is the attribute distance between records in the original dataset and anonymised dataset according to the normalised Euclidean Distance. The higher the error, the greater is the information loss. This metric can take values in the interval of [0,1] [12].
- *Non-uniform entropy:* This metric tries to evaluate and quantify the differences within attribute value distributions. To calculate the non-uniform entropy for a transformed dataset, the non-uniform entropy of each column

[3] https://arx.deidentifier.org/
[4] https://arx.deidentifier.org/overview/metrics-for-information-loss/

has to calculated and summed up. Non-uniform entropy compares the frequency of each feature value in the original dataset and the transformed dataset. This basic idea does not work well for local recording. Therefore, this utility metric will be calculated as follows: First, the generalisation level for each record will be calculated, which is followed by identifying the records that are affected by that generalisation level. Finally, the information loss according to non-uniform entropy will be calculated for each generalisation level. Additionally, the calculated value will be scaled into the interval [0,1] [13].

- *Granularity:* This utility metric captures the granularity of the data. For numerical attributes, the granularity of the generalisation intervals will be determined by the possible interval end points created during the discretisation. This metric can take values between 0 and 1. [14].

Measuring the effectiveness of anonymised data via statistical analysis tasks, such as a predictive machine learning model, is investigated e.g. in [15]. The authors compare applying six different algorithms, with very diverse results. The authors only evaluated the setting of 2-anonymous datasets, which would generally be regarded as too low.

A scheme for controlling over-generalisation of less identity-vulnerable QIs in diverse classes by determining the importance of QIs is presented in [16]. Comparing this scheme to others (such as Mondrian[17]), the authors measure accuracy on Decision Trees, Random Forests and SVMs. Their performance on large factors of k not only remains stable, but in some cases increases with k.

Effects of suppressing records costly to anonymise, instead of generalising several other records as well, has been studied in [18], on a number of binary classification problems. Multi-class problems are addressed in [11], with a focus of selectively deleting outliers to reduce the information loss during the anonymisation process. The authors consider Logistic Regression, SVMs with linear kernel, Random Forest, as well as Gradient Boosting.

3 Methodology

Data In our experiments, we use the Adult Data Set[5] from the UCI Machine Learning Repository. The dataset is prepared with the same steps as described in [11]. The dataset contains some missing values which will be eliminated due to its small number and therefore the dataset has 30162 data entries. The dataset has 15 attributes, only 14 of them will be used for the experiments since the attribute "education" represents the same information as "education-num", only differently encoded. To ensure a proper distribution of each attribute, we modify the column "native-country" to only contain *US-States* and *Non-US* since the the value *US-States* dominated the attribute distribution over 90% over all other countries.

[5] https://archive.ics.uci.edu/ml/datasets/Adult

Contrary to the default task in this dataset, which is a binary prediction of the income level, we evaluate a more challenging multi-class task. To this end, we define two different target variables, "education" and "marital-status". For "education" we group the 16 continuous "education" levels into four groups, while for "marital-status" we leave the dataset unmodified.

k-anonymity We utilise the Flash and SaNGreeA algorithms described above, which use global and local generalisation, respectively. For Flash, to evaluate the effectiveness of different candidate datasets, we created a multitude of these datasets, namely the ten best, one from the middle of the solution space and the worst found solution after the anonymisation process for a given k. We produce perturbations with $k = 3, 7, 11, 15, 19, 23, 27, 31, 35, 100$, and compare with the original, unmodified dataset.

Classification In order to measure the quality of the anonymised datasets for practical use, we train multiple classification algorithms with the dataset. We use Gradient Boosting, Random Forest, Logistic Regression and Linear SVC as classification algorithms of the python scikit-learn framework [6]. To avoid any optimisation bias towards a specific dataset, only a limited hyper-parameter optimisation has been conducted.

We primarily use the F1 score as the evaluation metric in our experiments. F1 measures the test's accuracy by taking both recall and precision into account. All exported anonymised dataset will be executed with the defined machine learning pipeline.

Correlation Analysis Beside comparing classification results directly, this paper aims to find relationships between the classification results and the utility metrics which characterise the anonymised datasets. To this end, we calculate the correlation between F1 score and the utility measurements. Our method calculates correlation based on the Pearson correlation coefficient as implement in Python library *scipy*[7].

We compare the earlier mentioned objective function *Loss*, *Record-level squared error*, *Non-uniform entropy*, and *Granularity*.

4 Evaluation and Analysis

In this section, we describe and discuss our experiments. We start with a general comparison of the effectiveness of the k-anonymous data, as seen in Figure 1.

We can see in all plots that there is a decline in classification effectiveness when anonymising the data, compared to the baseline of no anonymisation. However, there are very large differences in how the single classifiers are affected. For example, Logistic Regression immediately drops in F1 score when performing

[6] https://scikit-learn.org/stable/index.html

[7] `https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pearsonr.html`

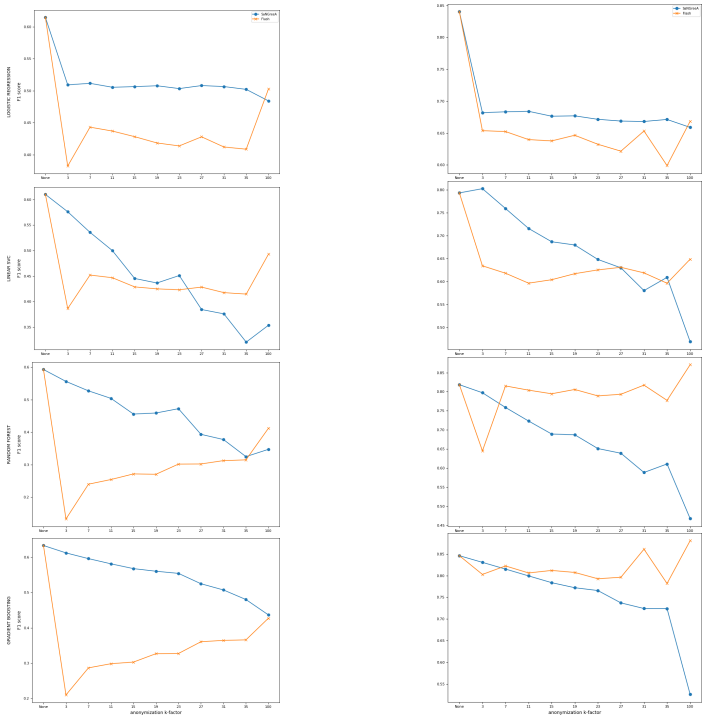


Fig. 1: Classification results for Flash and SaNGreeA, for several k values and two classification targets. (left: *education*, right: *marital status*

an anonymisation of $k = 3$. However, for higher values of k, the further deterioration is rather low. It can be observed that the local generalisation provided by SaNGreeA performs better in this case. A similar observation can be made by the Linear Support Vector classifier (SVC), which is not surprising, as these two classification models have a rather similar objective function they minimise. However, for SVC increasing k, and at some point, the global generalisation of Flash becomes superior.

For the ensemble methods of Random Forests and Gradient Boosting, the results are somewhat different. In general, there is a large deterioration in effectiveness between the baseline and $k = 3$, however the effectiveness appears to increase for larger values of k. For the target "marital-status", the global anonymisation of Flash is performing better for these classifiers. Especially for the bagging method of Random Forests for the same target, the drop in effectiveness is relatively low.

As a conclusion, the overall performance of the ensembles on the anonymised data is at a comparable level to the unabridged data, even for relatively large values of k.

We now specifically analyse the difference in utility when not using just the best, but also multiple solutions found by the Flash algorithm.

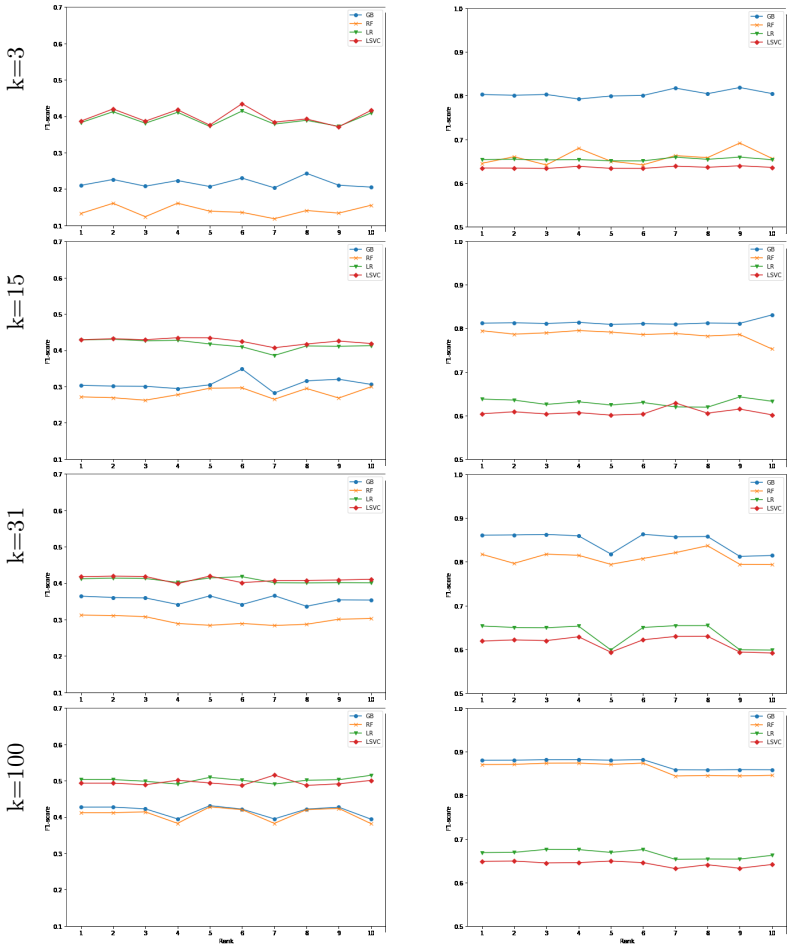


Fig. 2: F1 scores for the best ten datasets (left: education; right: marital status)

Figure 2 shows the classification results measured by the F1-score of the best ten output dataset for each k value and classification method, for "marital-status" and "education", respectively. In general, the fluctuations on F1-score are rather minute. It is visible from the plots that there are no significant differences between the classification results along the best ten datasets. While for some values of k, there is a slight decrease in the classification effectiveness (e.g. $k = 100$ for marital status), for other values, such as $k = 31$ on the marital status target, the tenth solution is actually the best.

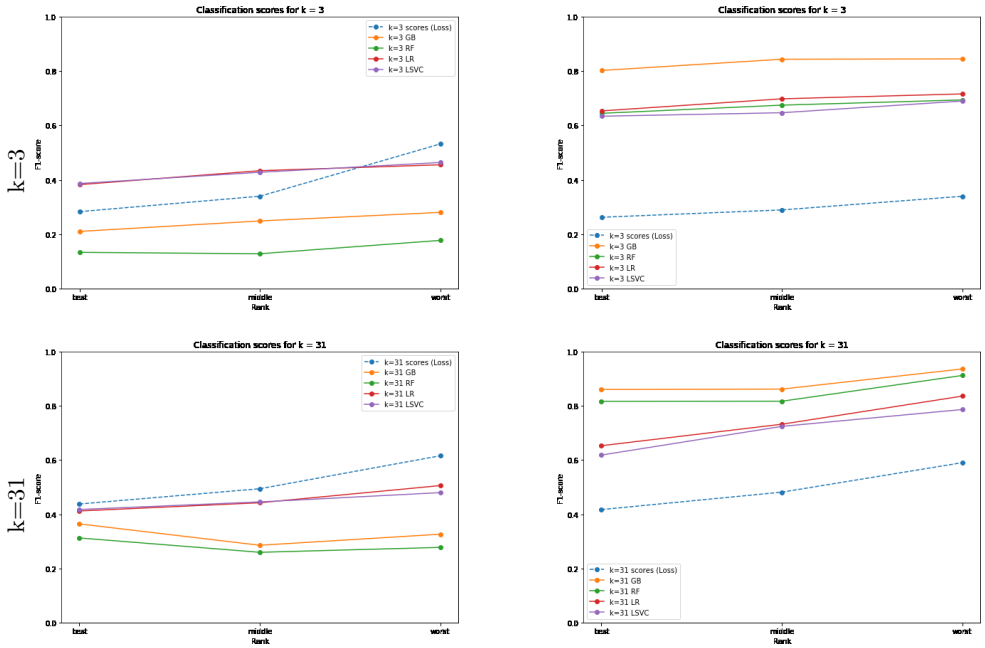


Fig. 3: Classification results of the best, middle and worst found dataset of the solution space

In order to investigate a wide spectrum of the solution space we also analysed the worst found dataset for each k, and one from the middle. Figures 11 and 12 show the overall rankings for these datasets (i.e. they indicate how many different solutions were found). The left column of Figures 3 and 8 shows the classification results for the best, middle and worst found dataset in the solution space for "marital-status", whereas the right column for "education". As we can see, there are no significant degradation in the classification performances along these datasets. Moreover, in some cases we found a better classification performance for the "worst" dataset than for the best. The dashed blue line on each diagram shows the objective function score results for each dataset.

The left columns of Figures 4, 5, 9 and 10 show the investigated utility metrics for each k value. On the right side of the figures, we see the correlation results between F1 score and Loss for each k value and each classifier; since Loss was the objective function, we examined the correlation of this value. The scores are computed via the Pearson correlation coefficient, where 1 means strong relationship, while -1 means negative correlation. In order to find reasonable correlations, we multiplied the correlation value by -1, since if Loss is higher, we expect worse classification results. As we can see, there is no clear relation between the classification results (F1 score) and the Loss score for "education". For $k = 100$, we observe a clear trend and moderate correlation, and further for

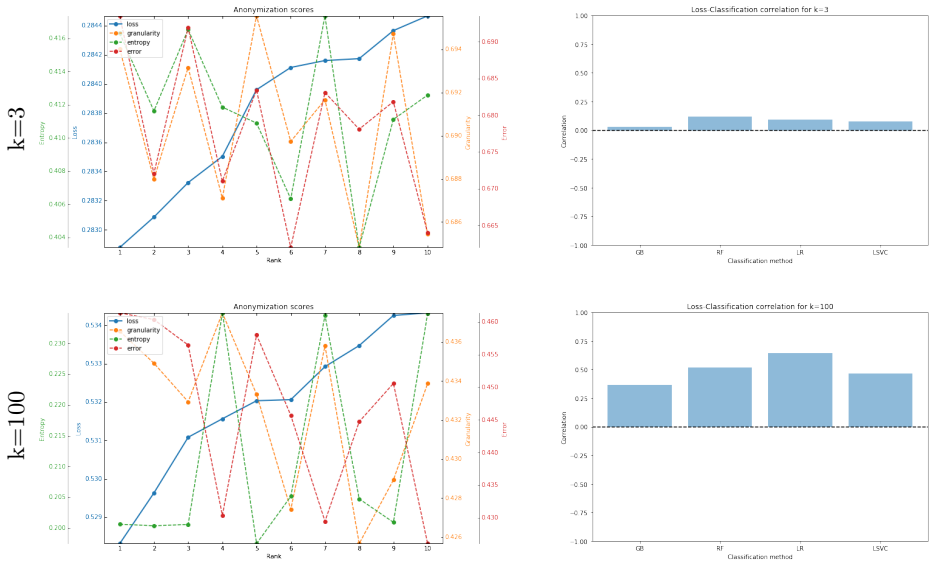


Fig. 4: Utility metrics and loss correlation for *education*

the linear models (Logistic Regression and linear SVC) for $k = 31$. For "marital-status", we can observe an overall correlation between all metrics and the F1 score for "marital-status" with $k = 3$, and to some extent also fo $k = 100$. There's an overall indirect correlation, though not that strong, for "marital-status" and $k = 1$. The other settings show either none, or no clear trend of correlation.

In order to investigate not only the Loss but also other utility metrics, we correlated the classification results (F1 score) with all scores. Figure 7 shows the correlation results for "marital-status", and Figure 6 for "educaton". As the plots show, we cannot derive global rules for the correlation. However, we can observe some case specific strong correlations. The record-level squared error and the non-uniform entropy correlates strongly for $k = 3$ on the "education" target attribute with the classification results, while granularity shows also a strong relationship for $k = 31$ and $k = 100$ on the same target variable. Loss also correlates for $k = 15$ on "marital-status" and some on "education". We further observe correlation for Logistic Regression and Linear SVC for $k = 31$.

5 Conclusion and Future Work

In this paper, we performed an analysis on the utility of k-anonymous datasets for specific classification tasks. We investigated (i) the differences between the multiple (syntactically valid) solutions found by heuristic anonymisation techniques, considering the ten best, as well as one from the middle of the solution space and the worst generated dataset. We can conclude that there is very little

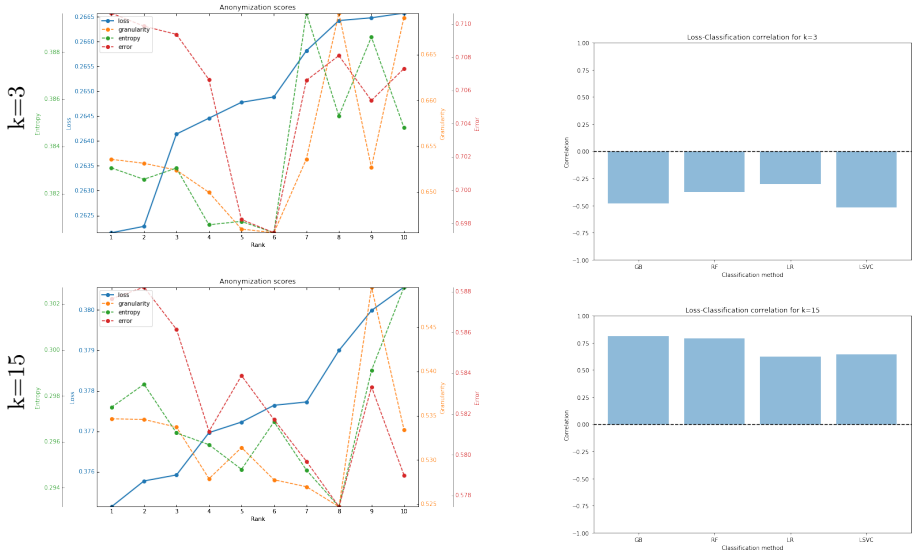


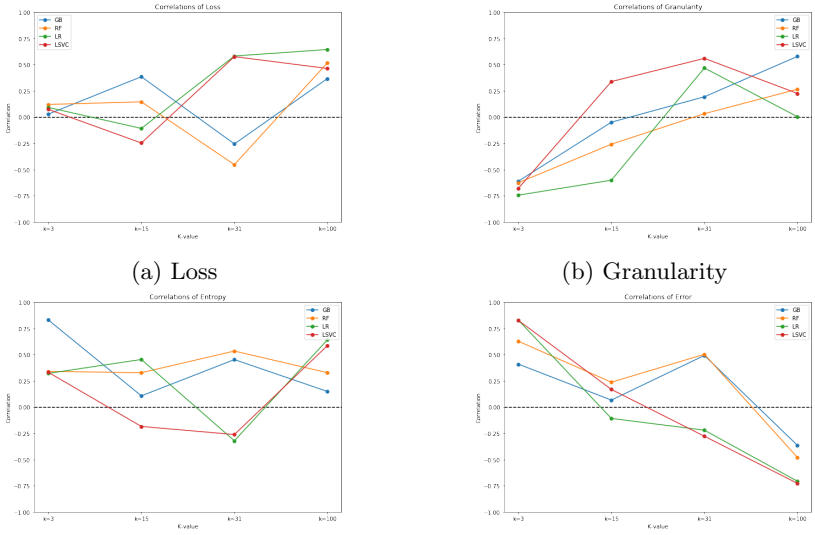
Fig. 5: Utility metrics and loss correlation for *marital-status*

difference in these solutions, which entails that the effectiveness of the resulting dataset is rather stable, and not influenced by potentially minute aspects in the heuristic. In some cases, even the supposedly worse solutions marginally outperform the best solution. We further investigated (ii) whether there is a correlation between measure that estimate the data utility directly on the dataset, versus the utility for the specific classification task. We specifically analysed Loss, Granularity, Non-uniform Entropy and Record-level squared error. Although, we could not derive any global rule of these correlations that can be applied independently of the task or the k value, we could see some specific correlations between classification and utility metrics. We can conclude that there is no overall, reliable correlation between these two measures, and it is thus not generally possible to estimate the classification performance based on the measures from the dataset alone.

Future work will focus on extending this analysis to further machine learning tasks such as regression, and will include further datasets. We will also extend the analysis to multiple solvers of the k-anonymity problem.

Acknowledgements

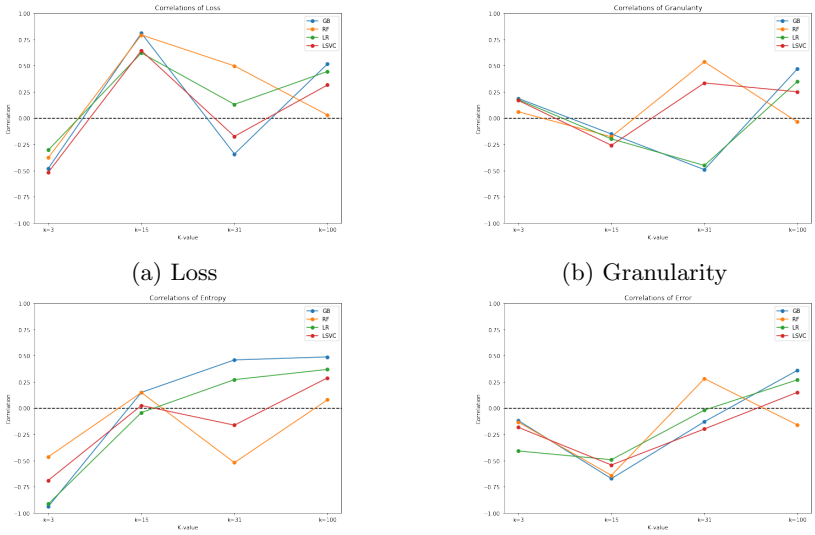
This work was partially funded by the BRIDGE 1 programme (No 871267, "WellFort") of the Austrian Research Promotion Agency (FFG), the EU Horizon 2020 research and innovation programme under grant agreement No. 826078 (Project "FeatureCloud"). SBA Research (SBA-K1) is funded within the framework of COMET — Competence Centers for Excellent Technologies by BMVIT, BMDW, and the federal state of Vienna, managed by the FFG.



(a) Loss (b) Granularity

(c) Entropy (d) Error

Fig. 6: Correlation results for all investigated utility metrics for *education*



(a) Loss (b) Granularity

(c) Entropy (d) Error

Fig. 7: Correlation results for all investigated utility metrics for *marital-status*

Appendix

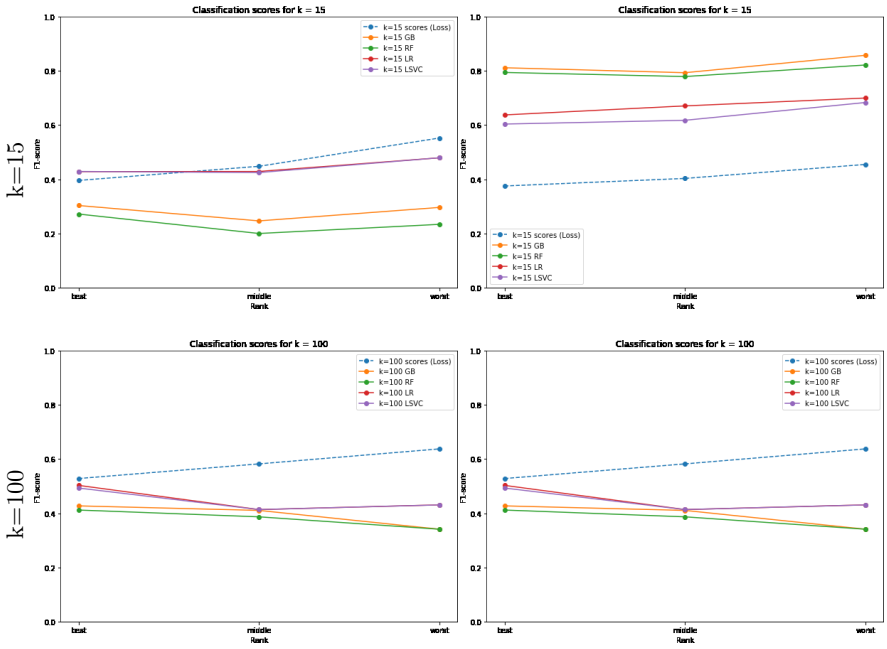


Fig. 8: Classification results of the best, middle and worst found dataset of the solution space

k value	middle rank	worst rank
3	46	92
15	48	96
31	40	80
100	16	33

Fig. 11: Rankings for *education*

k value	middle rank	worst rank
3	36	72
15	36	72
31	40	81
100	33	66

Fig. 12: Rankings for *marital-status*

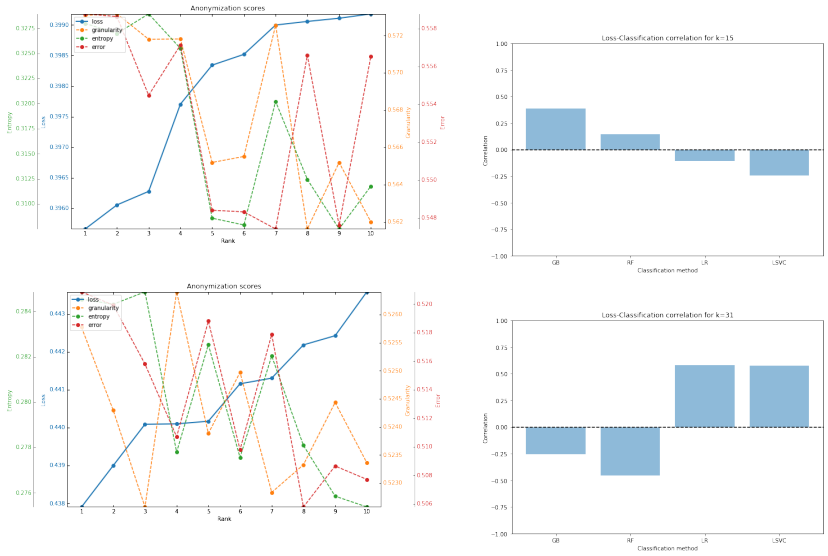


Fig. 9: Utility metrics and loss correlation for *education*

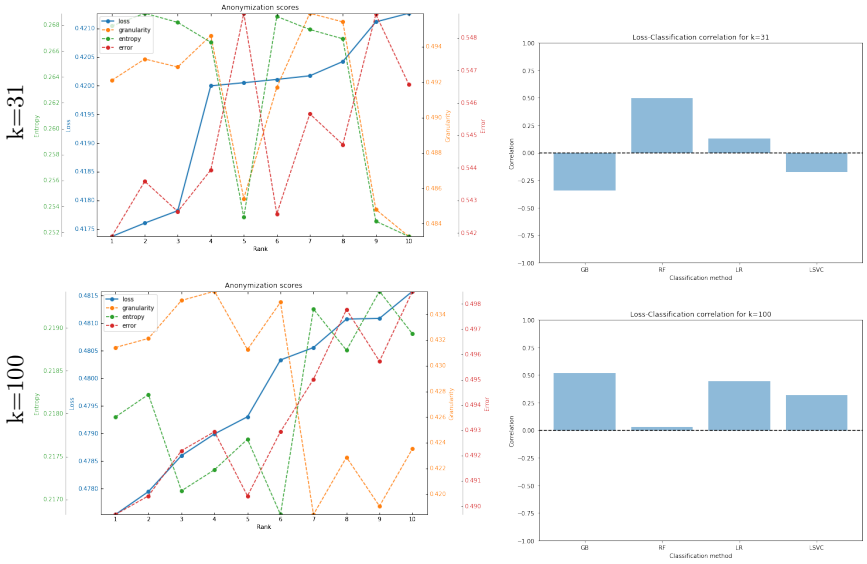


Fig. 10: Utility metrics and loss correlation for *marital-status*

References

1. Cynthia Dwork. Differential Privacy. In *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
2. Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In *Inter. Conference on Availability, Reliability and Security (ARES)*, Canterbury, UK, 2019. ACM.
3. Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, and Ashwin Machanavajjhala. Privacy-Preserving Data Publishing. *Foundations and Trends in Databases*, 2(1-2), 2009.
4. Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, 1998.
5. Latanya Sweeney. K-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 2002.
6. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, 2006.
7. Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, 2007. IEEE.
8. Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
9. Florian Kohlmayer, Fabian Prasser, Claudia Eckert, Alfons Kemper, and Klaus A. Kuhn. Flash: Efficient, Stable and Optimal K-Anonymity. In *International Conference on Privacy, Security, Risk and Trust and International Confernece on Social Computing*, Amsterdam, Netherlands, 2012. IEEE.
10. Alina Campan and Traian Marius Truta. Data and Structural k-Anonymity in Social Networks. In *Privacy, Security, and Trust in KDD*, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
11. Bernd Malle, Peter Kieseberg, and Andreas Holzinger. DO NOT DISTURB? Classifier Behavior on Perturbed Datasets. In *Machine Learning and Knowledge Extraction*, volume 10410, Cham, 2017. Springer International Publishing.
12. Jordi Soria-Comas, Josep Domingo-Ferrer, David Sanchez, and Sergio Martinez. t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation. *IEEE Transactions on Knowledge and Data Engineering*, 27(11), 2015.
13. Fabian Prasser, Raffael Bild, and Klaus A. Kuhn. A Generic Method for Assessing the Quality of De-Identified Health Data. *Studies in Health Technology and Informatics*, 228, 2016.
14. Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Edmonton, Alberta, Canada, 2002. ACM Press.
15. Hayden Wimmer and Loreen Powell. A comparison of the effects of k-anonymity on machine learning algorithms. In *Proceedings of the Conference for Information Systems Applied Research*, 2014.
16. Abdul Majeed, Farman Ullah, and Sungchang Lee. Vulnerability- and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data. *Sensors*, 17(5), 2017.
17. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. Mondrian Multidimensional K-Anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, Atlanta, GA, USA, 2006. IEEE.

18. Bernd Malle, Peter Kieseberg, Edgar Weippl, and Andreas Holzinger. The Right to Be Forgotten: Towards Machine Learning on Perturbed Knowledge Bases. In *International Conference on Availability, Reliability, and Security*, Cham, 2016. Springer International Publishing.