Poster: The impact of data sampling in the anonymization pipeline

Jenno Verdonck, Kevin De Boeck, Michiel Willocx, Jorn Lapon, Vincent Naessens
imec-DistriNet
KU Leuven
Ghent, Belgium
firsname.lastname@kuleuven.be

***Abstract*—An increasing number of companies are selling data as an additional source of revenue, or acquire data from other parties to optimize their business. In many cases, the shared data contains sensitive personal records. According to the GDPR regulation, personal data should be anonymized before it is released to third parties. A frequently applied technique is the k-anonymity metric, which ensures that every record in the dataset becomes indistinguishable from K other records through data generalization. This work combines generalization techniques with sampling. By adding a sampling step in the anonymization pipeline, additional uncertainty is introduced towards a potential attacker. As attackers can no longer be sure that an individual is in the sampled dataset, the re-identification risk is mitigated. This work proposes and evaluates multiple sampling techniques. Both the privacy and the utility properties of the anonymized datasets are embraced. The utility of the anonymized datasets is further evaluated in a machine learning use-case.**

***Index Terms*—anonymization, sampling, privacy, utility**

1. Introduction

Data collection and processing have become aspects of increasing importance in the daily operation of many businesses and organizations. Amongst others, data is employed for optimizing production processes and to increase the effectiveness of marketing campaigns. Hence, sharing (or trading) data can be very lucrative, and might even bootstrap cooperation between organizations. While data sharing exposes great opportunities for companies, caution should be taken. First, the European GDPR regulation states that datasets containing personal information should be anonymized before they are released. This means that a record in the shared dataset can no longer be linked to an individual afterwards. Second, thoughtless release can undermine the competitiveness of companies. Anonymization techniques can be applied to mitigate these threats.

An often cited and applied strategy for data anonymization relies on privacy metrics such as k-anonymity [1]. The goal of this metric is to generalize attribute values in such a way that each individual becomes indistinguishable from at least $k-1$ other individuals in the dataset. However, constructing a feasible anonymized dataset is no sinecure for most organizations. Some information is inevitably lost during anonymization caused by generalization. Hence, companies often struggle to find a satisfactory balance between the privacy and the utility of an anonymized dataset. Moreover, solely applying k-anonymity does not always protect against attackers with membership knowledge, a key assumption in the prosecutor attacker model [2]. This work argues that the aforementioned risk is mitigated by applying an additional sampling step in the anonymization pipeline. An attacker can no longer be sure that the target is in the dataset, making it harder to re-identify individuals. The prosecutor becomes a journalist, corresponding to the also well-known but less powerful journalist attacker model.

Contributions. This paper presents preliminary results of our research on combining traditional anonymization techniques (i.e. k-anonymity) and sampling. Three different sampling strategies are presented, implemented and assessed. The impact of each strategy on both the utility and the privacy properties of the remaining data is evaluated in a practical machine-learning use-case.

The remainder of this work is structured as follows. Section 2 points to related work. Section 3 describes different sampling strategies. Thereafter, Section 4 details the evaluation methodology. Lastly, Section 5 outlines the conclusions and points to future work.

2. Related work

Sampling has been the subject of previous research. Rocher et al. [3] demonstrate that solely applying random sampling does not sufficiently protect the privacy of individuals in the dataset. Hence, this work combines sampling with techniques for k-anonymity. Other research shows that random sampling in combination with k-anonymity can achieve differential privacy [4], [5]. Their main focus is redefining differential privacy in this context. Our work is complementary as it considers multiple sampling strategies and focuses on the utility-privacy balance. Still others do focus on the privacy-utility balance. [6] and [7] assess this balance in an optimization and an association rule mining case respectively. They do however solely focus on anonymization metrics and do not embrace sampling. Previous work [8] combined k-anonymity and sampling, and focused on the privacy-utility balance in an optimization use-case. This paper expands on this work by assessing this balance in a machine-learning setting, and by applying multiple sampling strategies. When training machine learning models on anonymized data, additional steps are required in order to apply the machine

learning models with non-anonymized data. Inan et al [9] elaborate on this challenge.

3. Sampling strategies

This work combines sampling with attribute generalization (i.e. k-anonymity). In the anonymization pipeline, the sampling step can be executed either before (*pre-sampling*) or after (*post-sampling*) the generalization step. Both approaches have advantages. *Pre-sampling* reduces the size of the dataset, and hence the complexity of the generalization step. *Post-sampling* enables more intelligent sampling based on the output of the generalization algorithm. Note that – especially in larger datasets – both strategies can be combined.

This work evaluates and compares three sampling strategies, namely *random sampling*, *stratified sampling* and *balanced stratified sampling*. Note that the latter two are only applied in *post-sampling*, as they require the output of the generalization algorithms (i.e. the equivalence classes of size $\geq k$):

- **Random sampling.** The main advantage of this straightforward sampling technique is that, by definition, no bias is introduced. Nevertheless, this technique has one major disadvantage. Because the sampling is completely random, there are no guarantees that records are removed from each equivalence class. Therefore, some equivalence classes can remain complete, giving an advantage to attackers with membership knowledge of the original dataset.
- **Stratified sampling** tackles the aforementioned problem by forcing that records are removed from each equivalence class. This method first calculates the amount of records that need to be removed from each equivalence class (based on the sizes of the equivalence classes), after which the required amount of records are removed from each equivalence class randomly.
- **Balanced stratified sampling.** Data is increasingly acquired for use in machine learning applications. Many machine learning techniques heavily benefit from a balanced target attribute. However, by nature, many datasets are unbalanced. The *balanced stratified sampling* is a variation on the *stratified sampling* technique. Instead of applying full random sampling within each equivalence class, priority is given to removing a record that is over-represented in the target attribute.

4. Evaluation methodology

The goal of this work is to assess the impact of anonymization on the utility of datasets when applied in machine learning. Fig. 1 illustrates the evaluation methodology. The original dataset is first anonymized in a two-step process, combining sampling (*pre- or post-sampling*) and generalization (for achieving k-anonymity). Thereafter, a machine learning model is created using the gradient boosted decision tree algorithm by scikit-learn. Finally, the privacy and utility of the anonymized dataset are evaluated. The experiments are executed on census data extracted by the *Folktables* tool [10]. The dataset of 1.6M records contains attributes such as age, place of birth, sex, marital status and income. The goal is to create an accurate machine learning model to predict whether the income of citizens reaches a certain threshold. The tests are repeated for a set of different values for k, sampling strategies and sample sizes. Each test is executed multiple times in order to rule out sample-dependent results. The remainder of this Section first outlines both the utility and privacy measurements used for the evaluation, after which some preliminary results are presented and discussed.

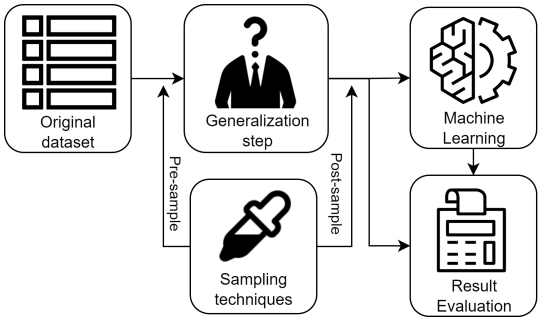


Figure 1. Evaluation workflow

4.1. Privacy measurements

The privacy properties of the anonymized datasets are evaluated by means of two metrics, namely *re-identification risk* and *certainty*. For calculating the risks, the *journalist risk* formula is applied [2], as the sampling step during anonymization ensures that an adversary is uncertain about the presence of the target in the anonymized dataset. The formula for calculating the journalist risk is presented in Equation (1), with f_j and F_j being the sizes of the equivalence classes in the sampled and the non-sampled dataset respectively. Note that the journalist risk is independent of the sample size. The *certainty* is calculated using Equation (2). It represents the certainty of an attacker that the target is in the sampled dataset.

$$Risk_j = \frac{f_j}{F_j} \cdot \frac{1}{f_j} = \frac{1}{F_j} \quad (1) \qquad Certainty_j = \frac{f_j}{F_j} \quad (2)$$

4.2. Utility measurements

This work focuses on two aspects of utility, namely case-agnostic utility and utility when applied for a specific machine learning purpose. In order to assess the former, this work analyses the selected generalization levels. Harsher generalizations typically imply increased information loss. It also compares the distribution (*chi-square* and *F-test*) of the data in the original dataset to the anonymized dataset. The machine learning utility is measured using the $accuracy$ (3) and F_1-$score$ (4). The models are evaluated by applying a 5-fold cross-validation on the dataset. In the equations below, TP and TN represent the amount of true positive and true negative predictions respectively, FN represents the amount of false negatives.

$$accuracy = \frac{TP + TN}{\#records} \quad (3)$$

$$F_1\text{-}score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (4)$$

4.3. Preliminary results

While this research is ongoing, some preliminary results are already available.

Fig. 2 presents a violin plot of the (journalist) risk for random pre- and post-sampling with k=20 and sample size 1/8. First of all, this plot demonstrates that results can vary between pre- and post-sampling. This difference can be attributed to the fact that pre-sampling reduces the amount of records in the dataset, and thereby the likeliness that records can be grouped in groups of size k. Therefore, the generalization algorithm is forced to impose harsher generalization levels to reach k-anonymity. Because the pre-sample has undergone harsher generalizations, the risk is lower compared to the post-sample. However, it should be noted that the post-sample dataset will score higher in utility metrics.

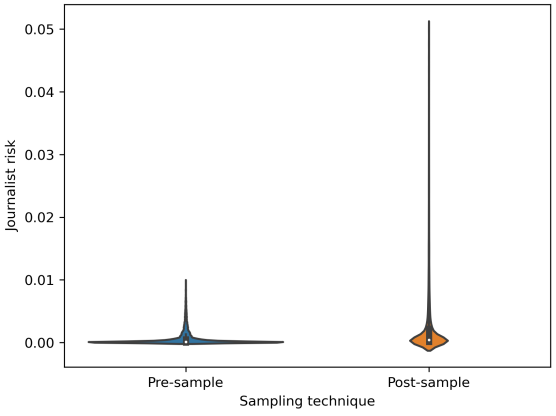


Figure 2. Journalist risk violin plots for random pre- and post-sampling with k = 20, samplesize = $\frac{1}{8}$

Fig. 3 presents a violin plot of the *certainty* for both the *full random* and *stratified* post-sample technique (k=5, sample size 0.5). This figure demonstrates that, while for most records the *random sampling* achieves its goal of introducing uncertainty for attackers, a set of records is underaffected by the sampling step (all records >0.5 in the left graph). Almost 50% of the records have a certainty over 0.5 and 0.2% (i.e. more than 2K records) have a certainty of over 80%. The *stratified sampling* strategy on the other hand boasts a maximal certainty of 0.5 (= the sample size).

5. Future work and conclusions

This work presented an overview of work in progress research on the effects of applying an additional sampling step in an anonymization pipeline. Different sampling strategies are presented and evaluated with respect to privacy and utility. However, many more experiments are required to finish this work. Firstly, we suspect that different combinations of sampling and generalization levels will achieve similar privacy properties but varying utility levels. Completing our experiments will lead to guidelines

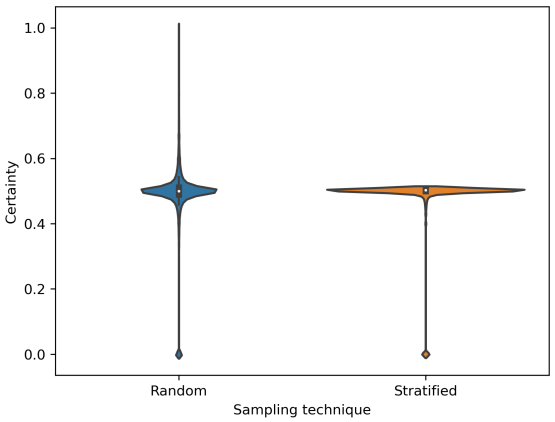


Figure 3. Certainty violin plots for random and stratified post-sampling with k = 5 and samplesize = $\frac{1}{2}$

for achieving optimal utility with respect to a certain privacy level. Moreover, based on our initial tests, our privacy and utility measurement metrics will undergo fine-tuning. Lastly, our tests currently cover one dataset and one specific machine learning problem. Expanding these tests to other data and algorithms will undoubtedly provide us with more complete and generic results.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[2] K. El Emam, *Guide to the de-identification of personal health information*. CRC Press, 2013.

[3] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[4] R. Bild, K. A. Kuhn, and F. Prasser, "Safepub: A truthful data anonymization algorithm with strong privacy guarantees," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 1, pp. 67–87, 2018. [Online]. Available: https://doi.org/10.1515/popets-2018-0004

[5] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy," in *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, 2012, pp. 32–33.

[6] R. Hoogervorst, Y. Zhang, G. Tillem, Z. Erkin, and S. Verwer, "Solving bin-packing problems under privacy preservation: Possibilities and trade-offs," *Information Sciences*, vol. 500, pp. 203–216, 2019.

[7] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 517–526.

[8] K. De Boeck, J. Verdonck, M. Willocx, J. Lapon, and V. Naessens, "Dataset anonymization with purpose: a resource allocation use case," in *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*. IEEE, 2021, pp. 202–210.

[9] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in *2009 IEEE 25th International Conference on Data Engineering*. IEEE, 2009, pp. 429–440.

[10] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," *arXiv preprint arXiv:2108.04884*, 2021.

KU LEUVEN

DEPARTMENT OF COMPUTER SCIENCE

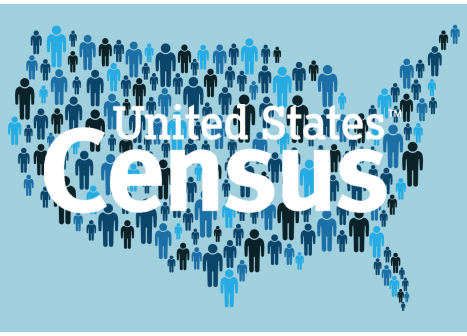
The impact of data sampling in the anonymization pipeline

Jenno Verdonck, Kevin De Boeck, Michiel Willocx, Jorn Lapon, Vincent Naessens
[1]imec-DistriNet, KU Leuven

Abstract

An increasing number of companies are selling data as an additional source of revenue, or acquire data from other parties to optimize their business. In many cases, the shared data contains sensitive personal records. According to the GDPR regulation, personal data should be anonymized before it is released to third parties. A frequently applied technique is the k-anonymity metric, which ensures that every record in the dataset becomes indistinguishable from K other records through data generalization. This work combines generalization techniques with sampling. By adding a sampling step in the anonymization pipeline, additional uncertainty is introduced towards a potential attacker. As attackers can no longer be sure that an individual is in the sampled dataset, the re-identification risk is mitigated. This work proposes and evaluates multiple sampling techniques. Both the privacy and the utility properties of the anonymized datasets are embraced. The utility of the anonymized datasets is further evaluated in a machine learning use-case.

Research methodology



Quasi-identifier			Sensitive
Gender	Age	Nationality	Income
Male	28	Russia	<50K
Male	29	Russia	<50K
Female	25	Japan	>50K
Male	41	US	<50K
Female	42	Canada	>50K
Female	44	US	<50K
Female	37	India	<50K
Male	39	Russia	<50K
Male	36	Japan	>50K

Original dataset

Data collection

Generalization
→ Individual privacy

Quasi-identifier			Sensitive
Gender	Age	Nationality	Income
Male	28	Russia	<50K
Male	29	Russia	<50K
Female	25	Japan	>50K
Male	41	US	<50K
Female	42	Canada	>50K
Female	44	US	<50K
Female	37	India	<50K
Male	39	Russia	<50K
Male	36	Japan	>50K

Quasi-identifier			Sensitive
Gender	Age	Nationality	Income
*	[25-30[	Asia	<50K
*	[25-30[	Asia	<50K
*	[25-30[	Asia	>50K
*	[40-45[	N-America	<50K
*	[40-45[	N-America	>50K
*	[40-45[	N-America	<50K
*	[35-40[	Asia	<50K
*	[35-40[	Asia	<50K
*	[35-40[	Asia	>50K

K-anonymity (ARX)

Sampling
→ Membership privacy

Quasi-identifier			Sensitive
Gender	Age	Nationality	Salary
Male	28	Russia	<50K
Male	29	Russia	<50K
Female	25	Japan	>50K
Male	41	US	<50K
Female	42	Canada	>50K
Female	44	US	<50K
Female	37	India	<50K
Male	39	Russia	<50K
Male	36	Japan	>50K

Quasi-identifier			Sensitive
Gender	Age	Nationality	Salary
Male	28	Russia	<50K
Female	25	Japan	>50K
Male	41	US	<50K
Female	37	India	<50K
Male	39	Russia	<50K

Random | Stratified | Balanced

Anonymization Pipeline

Evaluation

Privacy vs Utility balance

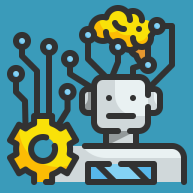
→ Risk
→ Certainty

→ F_1-Score
→ Generalization level

Quasi-identifier			Sensitive
Gender	Age	Nationality	Income
Male	[25-30[	Asia	<50K
Female	[25-30[	Asia	>50K
Male	[40-45[	N-America	<50K
Female	[35-40[	Asia	<50K
Male	[35-40[	Asia	>50K

Multiple anonymized versions of the dataset

Machine Learning



→ Income prediction model

Result evaluation

KU Leuven Technology Campus Ghent
Department of Computer Science
Gebroeders de Smetstraat 1
9000 Ghent, Belgium

DistriNet