A Data Utility-Driven Benchmark for De-identification Methods

Oleksandr Tomashchuk[1,2](✉), Dimitri Van Landuyt[2], Daniel Pletea[1],
Kim Wuyts[2], and Wouter Joosen[2]

[1] Philips Research, Eindhoven, Netherlands
{oleksandr.tomashchuk,daniel.pletea}@philips.com
[2] imec-DistriNet, KU Leuven, Leuven, Belgium
{dimitri.vanlanduyt,kim.wuyts,wouter.joosen}@cs.kuleuven.be

Abstract. De-identification is the process of removing the associations between data and identifying elements of individual data subjects. Its main purpose is to allow use of data while preserving the privacy of individual data subjects. It is thus an enabler for compliance with legal regulations such as the EU's General Data Protection Regulation. While many de-identification methods exist, the required knowledge regarding technical implications of different de-identification methods is largely missing. In this paper, we present a data utility-driven benchmark for different de-identification methods. The proposed solution systematically compares de-identification methods while considering their nature, context and de-identified data set goal in order to provide a combination of methods that satisfies privacy requirements while minimizing losses of data utility. The benchmark is validated in a prototype implementation which is applied to a real life data set.

Keywords: De-identification · Anonymisation · Pseudonymisation · Data utility · Privacy · GDPR

1 Introduction

The rise of the Internet-of-Things (IoT) and Big Data creates unprecedented opportunities to businesses, yet also the rapid development of these technologies raises many challenges. Increasing amounts of data are being entrusted to service providers, as big data analytics capabilities are increasingly powerful. However, new business and emerging models in the context of these technological advances also create much societal concern with respect to privacy and trust. As such, methods that improve protection of sensitive data for privacy purposes are rapidly gaining importance, and their proper usage is a success factor for digital business systems.

Many privacy-enhancing technologies (PETs) and privacy building blocks exist [34], and these vary in terms of complexity, practical applicability and architectural impact. Data de-identification is one sub-class of PETs that groups

S. Gritzalis et al. (Eds.): TrustBus 2019, LNCS 11711, pp. 63–77, 2019.
https://doi.org/10.1007/978-3-030-27813-7_5

methods which involve removing associations between the gathered or processed data and the identity of the data subject. This allows extensive use of data sets while preserving the privacy of individual data subjects. There are many methods for data set de-identification, varying from simply removing identifiable data to obfuscating and adding noise.

Despite existing survey efforts (e.g. [18]), the required knowledge on peculiarities of application of different de-identification methods to structured textual/numerical data is largely missing. From this perspective, we particularly focus on de-identification of such data. De-identification of unstructured data and multimedia data (audio/video/pictures) is out of the scope of this paper. Extensive review and performance evaluation of single- and multi-shot (image and video) re-identification algorithms can be found in [12], whereas a survey study of de-identification of multimedia content can be found in [30].

In the context of software engineering, sound approaches towards selecting appropriate de-identification methods are currently lacking. This lack of expertise can lead to re-identification attacks like the AOL and Netflix re-identification examples [28]. Employing different privacy models (e.g. k-anonymity [15], t-closeness [24], etc.) can mitigate such re-identification risks. However these privacy models have limitations that can be further addressed by combining them in the right way tailored per specific data sharing purpose. Considering the shortage of de-identification experts and the fact that de-identification highly depends on their expertise, not just systematization of methods is needed, but also automation of de-identification processes is desirable.

In this paper, we introduce a benchmark system for de-identification methods. It may be considered as an extension of approaches proposed by Xiong and Zhu [26] and Morton et al. [32]. This benchmark takes into account privacy requirements, unique properties of existing methods and data utility metrics, changes of data utility triggered by the de-identification processes, and goals of de-identified data. We present the design of the benchmark system, introduce its prototype implementation and validate it in the context of realistic data sets.

The remainder of this paper is structured as follows: Sect. 2 discusses the background of the paper, whereas Sect. 3 states the problem. Section 4 provides description of the proposed benchmark, Sect. 5 demonstrates the validation that we performed for our solution. Section 6 provides an overview of related work, Sect. 7 discusses important aspects of the benchmarking process. Section 8 concludes the paper and highlights future work.

2 Background

In general, de-identification is accomplished by removing the association between the (identifying) data and the data subject. It can be achieved by applying specific methods. Section 2.1 first establishes an overview of de-identification terminology. Then Sect. 2.2 lists existing de-identification approaches and summarizes existing support for de-identification.

2.1 Terminology and Concepts

There is unfortunately no single agreed-upon definition for privacy concepts. Several commonly-used definitions exist for the overarching concept of ‘privacy’, yet they all have a slightly different meaning [31].

Pfitzmann and Hansen [25] have defined anonymity of a subject as ‘*the subject is not identifiable within a set of subjects, the anonymity set*’. Sweeney [15] was first to specify the level of anonymity: k-anonymity of information means that the information for each person cannot be distinguished from at least k-1 individuals whose information is in the data set. The General Data Protection Regulation (GDPR) [33] defines anonymous information as ‘*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*’.

De-identification, anonymisation and pseudonymisation all describe the action of bringing data in an anonymous or pseudonymous state.

There is however also no consensus in terminology of, and relations between, ‘de-identification’, ‘anonymisation’ and ‘pseudonymisation’. To illustrate, there are noticeable differences in the definitions of anonymisation provided by ISO and NIST:

- ‘Process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party’ [2];
- ‘Process that removes the association between the identifying data set and the data subject’ [7];

Even more striking differences may be noticed with respect to definitions of de-identification provided by NIST [7] and Zuccona et al. [36]. Similar issues exist for definitions of pseudonymisation in the GDPR [33] and by NIST [7].

Terminological mismatch between different sources of information makes application of methods as well as understanding of fundamental concepts very challenging. In order to cope with it in this paper, we consider de-identification as a concept of a higher level, which covers both anonymisation and pseudonymisation, as follows:

- *De-identification* refers to any process of removing the association between a set of identifying data and the data subject [7];
- *Pseudonymisation* refers to processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person [33]; We also find pseudonymisation as a type of de-identification as claimed in [2];
- *Anonymisation* is a process through which personally identifiable information (PII) is irreversibly altered in such a way that a PII principal can no longer

be identified directly or indirectly, either by the PII controller alone or in collaboration with any other party [1].

2.2 De-identification Methods, Privacy Requirements and Loss of Data Utility

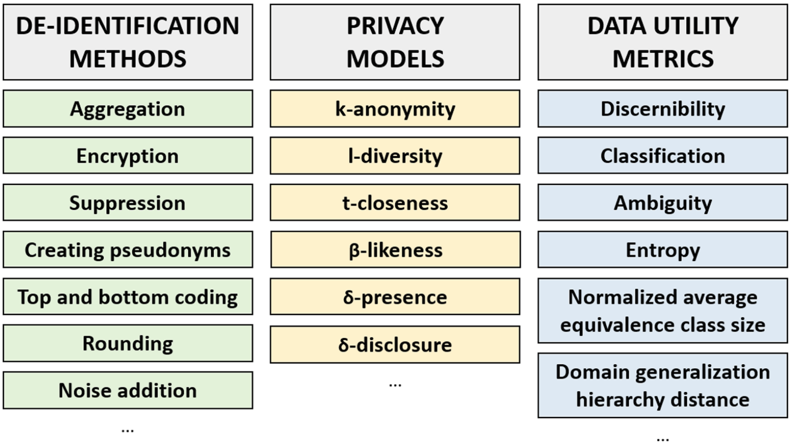
De-identification is a process implemented by de-identification methods, aimed at the realization of privacy requirements. Due to its transformative nature, this process always incurs losses in terms of data utility.

De-identification methods are used for reducing the amount of identifying information, and attaining the privacy requirements that are typically represented in privacy models that allow reasoning about the privacy of the data subject, and ensuring compliance. To this end, data utility metrics, underestimated in the current state-of-the-art, are very important for quantifying data utility losses and thus for assessing the level of de-identification, to create a de-identified data set suited for the intended use. Examples of state-of-the-art de-identification methods, data utility metrics, and privacy models are presented in Fig. 1.

Many de-identification methods are in existence and some attempts towards their systematization have already taken place. For example, Nelson lists existing methods in [23], but his systematization has a large degree of overlap between the de-identification methods. A more recent and far more precise systematization was introduced in the ISO 20889 standard [3]. It thoroughly describes de-identification methods that are grouped into eight distinct types, based on the nature of the techniques that they use. This systematization together with [6] and [13] forms a basis of a knowledge base that is required for selecting a suitable de-identification method.

Nevertheless, there is always a probability of having removed too much or too little information from the involved data set. In order to reduce this probability, one needs to have clearly-defined privacy requirements. In most cases, the de-identification need is triggered by risks, which are the basic drivers of privacy requirements. Transforming probabilistic risks into deterministic requirements is a challenge that can be addressed with the use of privacy models. Many privacy models are in existence, such as k-anonymity [15], l-diversity [19], t-closeness [24], β-likeness [10], δ-presence [21], δ-disclosure [11], etc. The selection of such a model depends on its applicability to particular cases.

However, knowledge of requirements and methods is hardly enough for proper de-identification. This is due to the fact that almost any data set contains potentially relevant information that can be used for gaining insights into the data, while the goal of de-identification is exactly to remove the link between the data subject and that information of interest. Such a removal is always accompanied by reduction in usefulness of the information. In order to take this into consideration, one needs to use data utility metrics that are capable of quantifying the changes in the utility of the information of interest. Many data utility metrics are in existence that are aimed at measuring different properties of a data set and a tailored selection of them as well as correct usage improves the results of the de-identification process [27].



DE-IDENTIFICATION METHODS	PRIVACY MODELS	DATA UTILITY METRICS
Aggregation	k-anonymity	Discernibility
Encryption	l-diversity	Classification
Suppression	t-closeness	Ambiguity
Creating pseudonyms	β-likeness	Entropy
Top and bottom coding	δ-presence	Normalized average equivalence class size
Rounding	δ-disclosure	Domain generalization hierarchy distance
Noise addition	...	...
...		

Fig. 1. Examples of state-of-the-art de-identification methods, privacy models, and data utility metrics

3 Problem Statement

As shown in the previous section, there is a broad variety in de-identification methods and heterogeneity in different tool implementations for each of these methods. If de-identification is not done properly, re-identification may take place which can lead to financial, reputational, and other kinds of losses. However, in practice, selecting the most suited method is a non-trivial task.

Firstly, the most suitable method depends highly on the application context: the nature of the data, privacy requirements, assumptions made about data accessibility, and potential attackers and their incentives all play a role.

Secondly, this selection process depends highly on the expertise and knowledge of the de-identification expert. For example, finding good parameters for these methods requires extensive know-how. In practice however, such expertise is not always readily available.

Thirdly, the selection and configuration of a method involves complex architectural trade-offs. As mentioned by Mittal et al. [9], a de-identification scheme can be evaluated from two complementary perspectives: data utility preservation and resistance to re-identification attacks. This is also corroborated by Lee et al. [17] through their demonstration of the relation between data utility and disclosure risk (depicted graphically in Fig. 2).

Considering these facts, an algorithm for defining the most suitable de-identification methods of data sets is necessary, because each data set, depending on the type of information that it contains and the availability of that information to attackers, needs a specific combination of de-identification methods for bringing the level of de-identification to a satisfactory level.

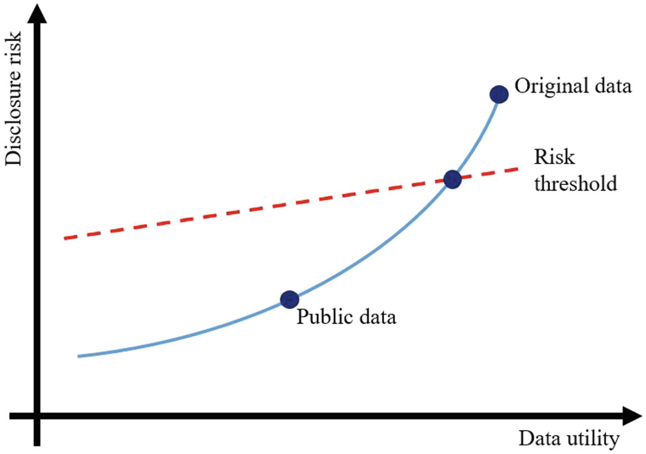


Fig. 2. Data utility vs. disclosure risk (from: [17])

4 Approach

As discussed in the previous section, selecting the most suited de-identification method for a specific application relies on multiple factors. In this paper, we present a systematic benchmark for de-identification methods. The benchmark system implements an exhaustive search for the most appropriate methods and combinations thereof in terms of two key factors: (i) adherence to the privacy requirements and (ii) data utility loss of the transformed data set.

The benchmark implements a two-phased approach: (i) in the expansion phase, candidate methods are generated, and (ii) in the reduction phase, these are filtered based on the privacy requirements (which act as a cut-off) and the data utility score.

The benchmark is exhaustive in the sense that it applies different methods at a fine-grained level (attribute-level) and allows combining different methods within the same data set. Furthermore, it systematically iterates over parameter values of the different methods, and thus allows assessing the impact of these parameters.

Figure 3 provides a graphical, flow-based representation of the proposed benchmark system, which in turn is refined throughout the following sections.

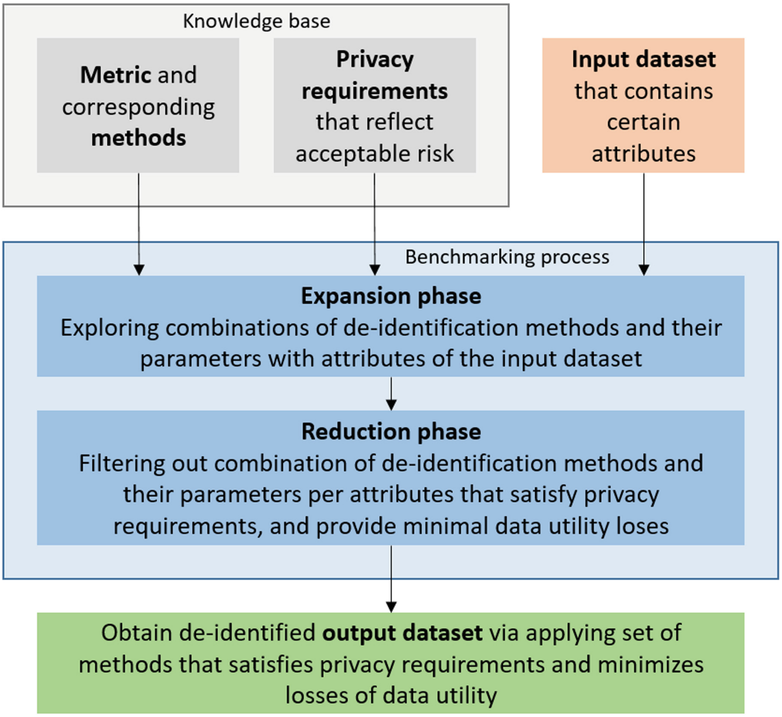

Knowledge base
Metric and corresponding methods
Privacy requirements that reflect acceptable risk
Input dataset that contains certain attributes
Benchmarking process
Expansion phase
Exploring combinations of de-identification methods and their parameters with attributes of the input dataset
Reduction phase
Filtering out combination of de-identification methods and their parameters per attributes that satisfy privacy requirements, and provide minimal data utility loses
Obtain de-identified output dataset via applying set of methods that satisfies privacy requirements and minimizes losses of data utility

Fig. 3. Flow diagram of the proposed approach

4.1 Input and Output

The input consists of a structured data set that has to be de-identified. We expect this data set to contain attributes that represent direct and/or indirect identifiers, and a finite amount of tuples. The data set may contain any kind of information that may be de-identified using the methods introduced in [3].

The output consists of the de-identified data set obtained using the de-identification methods that satisfy the privacy requirements and maximally preserve data utility. In general, it should not contain direct identifiers or information that is not relevant for reaching de-identified data set goals.

4.2 Knowledge Base

The knowledge base is a representation of information that is necessary for execution of the benchmark. However, the base should not be treated as a part of the proposed benchmark. Considering that the knowledge base does not belong to the processing part, inputs on it are neither exhaustive nor complete, and presented here to support gaining an understanding of the proposed approach.

Privacy Requirements. Privacy requirements represent the conditions that should be met by the de-identified data set. In general, they are met by selecting a certain privacy model (e.g. k-anonymity [15], l-diversity [19], etc.) and selecting a certain value as a threshold. For example, one can set a threshold by selecting $K = 20$ in case of using k-anonymity as a privacy model. The risk threshold is also represented in Fig. 2, which shows that failing to satisfy the threshold will lead to unwanted risk increase and probably various kinds of losses. Nevertheless, establishing such thresholds is a challenging task, as it also has to reflect the type of attackers, their capabilities, and the availability of data related to the data set. Profiles of attackers that may be useful for establishing privacy requirements can be found in [16].

Data Utility Metrics and Corresponding De-identification Methods. A metric represents a way of measuring the usefulness of the data in a certain context. In our case, it has to measure changes in data utility with regards to a specific goal. For example, if one is interested in obtaining information from the de-identified data set regarding the distribution of values for specific attributes, then the usage of the Discernibility Metric [14] may be reasonable, since it reflects the size of equivalence classes which have a direct link to the distribution of values. The choice of a specific metrics directly influences the selection of de-identification methods. Most of the methods are heterogeneous in nature, and thus a careful match that considers the nature of both metric and methods should take place before performing de-identification. It is of high importance to select and apply those methods, which led to changes in data utility that are measurable by selected metric. For example, the Normalized Average Equivalence Class Size Metric [14] can precisely measure changes of data utility caused by aggregation and rounding, but it becomes impossible to extract meaningful values of data utility change by applying it after suppression or replacement of values by pseudonyms. Another example that demonstrates the necessity of tailoring the selection of methods to the types of data is that if we have a data set with attribute "Address", applying permutation will not be helpful for improving local k-anonymity, but t/b (top and bottom) coding would be helpful.

4.3 Benchmarking Process

The simplified twofold representation of the benchmarking process in Fig. 3 highlights its two main phases: expansion and reduction. In practice, the border between them is not clearly delineated, and that is why we further refine these phases in the form of pseudo-code below. For demonstration purposes, we use k-anonymity as the main privacy model, but other models (l-diversity, t-closeness, etc.) may be used as well.

```
input  : privacy requirement $K$, privacy model $kan$, data utility metric
         $M$ and its loss implementation $dul$, de-identification methods
         $SM_i$ and their parameters $SP_{ij}$ suitable for given $M$, data set $D$
         that contains attributes $A_h$ which have to be de-identified;
output: combination $R_h$ of methods with specific parameters per every
         attribute that satisfy privacy requirements and minimize data
         utility losses;

foreach $A_h$ do
    foreach $SM_i$ do
        foreach $SP_{ij}$ do
            Obtain de-identified attribute $A'_h = SM_i(SP_{ij}, A_h)$;
            Obtain $PK = kan(A'_h)$;
            if $PK > K$ then compute data utility loss $dul(A'_h, A_h)$ and
            save it to a 3-dimensional buffer $BUF[h, i, j]$;
        end
    end
end
Obtain list $R_h \subset BUF$ that contains per every tuple $h$ indexes of
methods $i$ and parameters $j$, $dul$ of which is minimal;
```

5 Validation

As the most influential part of the proposed benchmark is the reduction phase, we decided to focus our validation explicitly on it. Considering its twofold nature, we have crystallized two fundamental concepts of the phase for validation purposes which can be found below:

1. Applying different methods to the very same piece of data leads to different results of de-identification;
2. Every de-identification method influences data utility in a unique way.

The first concept affects the part of the benchmark in which every possible combination of methods and parameters per attributes is applied and the resulting data set evaluated against the privacy requirements. The second concept in turn affects the part of the benchmark that involves searching for a combination of methods and parameters per attributes that minimizes data utility losses.

5.1 Validation of the First Concept

In order to validate this concept we created a Python script that under given privacy requirements based on k-anonymity delivers a result in form of a method, which, when applied, allows reaching the k-anonymity level closest to the target. Our script includes implementations of the following methods: shuffling (permutation), hashing, top and bottom coding, aggregation, and suppression. We applied the script to a Graduate Admissions data set [5] and fix parameters of the methods to ensure that changes in results can be attributed to the nature of the applied methods, instead of changed parameters. The validation results are shown in Table 1. These results demonstrate that for reaching different levels of k-anonymity, different methods should be used.

Table 1. Methods per attributes, which, when applied, lead to obtaining k-anonymity which is the closest to needed privacy level (K – the required level of k-anonymity)

Attribute	$K = 1$	$K = 5$	$K = 50$	$K = 300$
`Serial No.`	shuffling	aggregation	aggregation	suppression
`RE Score`	aggregation	t/b coding	t/b coding	t/b coding
`TOEFL Score`	suppression	suppression	suppression	suppression
`University Rating`	shuffling	shuffling	aggregation	aggregation
`SOP`	shuffling	shuffling	aggregation	aggregation
`LOR`	shuffling	shuffling	aggregation	aggregation
`CGPA`	suppression	suppression	aggregation	aggregation
`Research`	shuffling	shuffling	shuffling	aggregation
`Chance of Admission`	suppression	suppression	aggregation	aggregation

5.2 Validation of the Second Concept

In order to validate this concept, we created a Python script that compares data utility before and after applying the specified methods. The measurements are done with the usage of the following metrics: the Discernibility Metric, the Normalized Average Equivalence Class Size Metric, and the Probability Distribution Metric. As in the previous section, our script includes implementations of the following methods: t/b coding, aggregation, and rounding. Similarly to the approach that we used for the previous concept, we applied the script to the same data set and kept parameters of the methods constant in order to assure that changes in results originate from the nature of methods, but not from the changes of parameters.

The results are presented in Fig. 4, and they demonstrate that if we measure data utility with a certain metric after applying some de-identification methods, losses in data utility are quite different in every case. Even if the nature of metrics is quite similar (both Discernibility Metric and Normalized Average Equivalence Class Size Metric measure properties of equivalence classes), the pattern which can be observed on the corresponding graphs is not identical. Also, the results of applying the Probability Distribution Metric show that these methods in question do not perform well on attributes such as `Serial No.`, `University Rating`, and `SOP`, while being measured by the given metric.

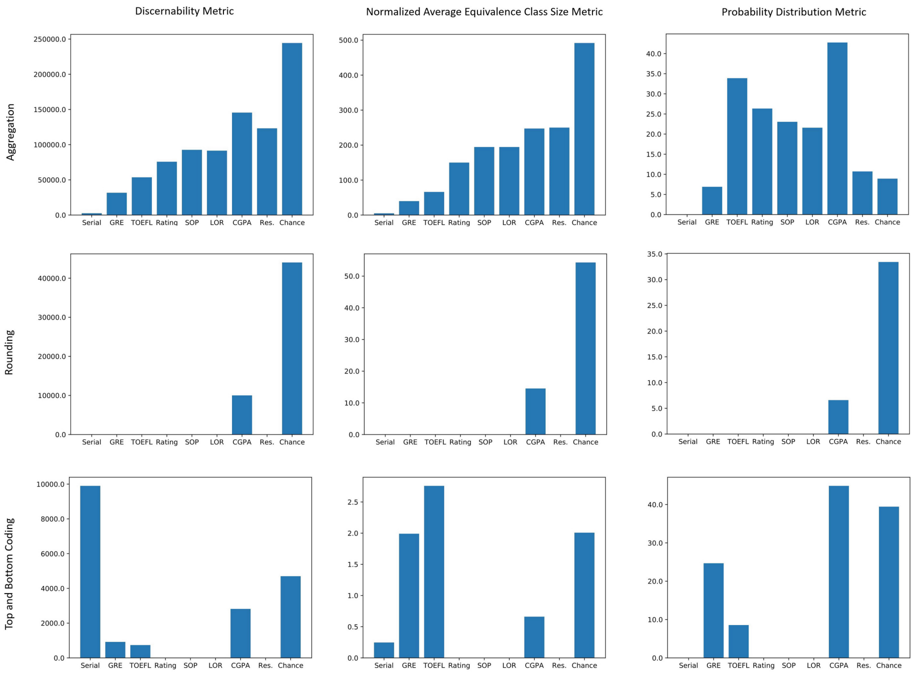


Fig. 4. Visualization of data utility losses per attribute under specific metric. The losses occur due to application of given de-identification methods. X axis represents the attributes of the data set. Y axis represent data utility loss (Discernibility Metric: min = 0, max = 250,000 (for given data set); Normalized Average Equivalence Class Size Metric: min = 0, max = 500 (for given data set); Probability Distribution Metric: min = 0%, max = 100%)

6 Related Work

Previous work in this field is mostly focused around utility and quality metrics. For example, LeFevre et al. [14] consider the Discernibility and Normalized Average Equivalence Class Size Metrics as measures of quality of anonymization in case of applying generalization or perturbation to the input data set. Goldberger and Tassa [8] contemplate loss, ambiguity, probability distribution, mutual information, and classification metrics as a way of measuring data utility. Also, there were some proposals for measuring data utility and level of de-identification through the application of game theory and entropy-based models [35]. However, these proposed metrics are relatively complex and applying them demands strong expertise in corresponding fields.

The importance of utility metrics is extensively discussed in [27]. Podgursky highlights that it is not clear whether a single metric which can accurately measure anonymization quality across all data sets and use-cases may exist, and that it is quite possible that a general metric will not accurately capture the data quality for a specific case. Also, he stresses that if any potential uses of a data set are known before anonymizing, it is generally advisable to tailor the quality metric on the expected use.

Templ and Meindl [20] complement this by stating that it is beneficial to evaluate the gain in explanation of parameters or variables when releasing de-identified data.

Xiong and Zhu in [26] tried to cope with the trade-off between data utility and privacy, but their approach is based only on information loss as a measure of data utility reduction, and data impurity as a measure of privacy gain. Ignoring the broader variety of existing metrics, methods, and privacy models puts restrictions on the added value of their work for practitioners and de-identification experts. Another problem of their contribution is that it does not consider the goal of the de-identified data set which is of high importance for performing proper de-identification. Similar consideration were also made in [4,22,29,32].

7 Discussion

A few aspects of the benchmark introduced in this paper are worth to be highlighted.

Firstly, tailoring of de-identification methods to data utility metrics is important, as this otherwise leads to situations in which methods reduce data utility significantly, but this is not properly highlighted during the reduction phase. As a result, the quality of the output may suffer significantly.

Secondly, it supports the use of multiple metrics simultaneously. This allows the usage of the benchmark when the utility of the information of interest in the de-identified data set may be quantified with different metrics. The output of the benchmark is a combination of methods that satisfies privacy requirements and minimizes losses of data utility, but it is possible to provide a ranking. This may be useful for cases when requirements are not strict, and allows exploration of these intermediate results for finding the optimal solution. This is due to the fact that slight strengthening of some requirements may lead to huge data utility losses or vice versa. While it makes the computations heavier, it may bring added value given that the relations between privacy requirements and the benchmark results are not linear in most cases.

Thirdly, privacy requirements, metrics and methods do not reflect the presence of direct identifiers, and so we expected them to be removed during the process of de-identification by other means.

Fourthly, the proposed benchmark is expected to be applied to real life IoT solutions during the process of their development. However, two requirements need to be satisfied for applying the benchmark: (i) it is necessary to have a dataset that is equal to the one that has to be de-identified at a predefined point of the solution's architecture, and (ii) the element that will be responsible for applying de-identification methods should have enough computational power for executing them. In case these requirements are met, the output of the benchmark will enable extension of the element's software by the most suitable de-identification methods.

8 Conclusion and Future Work

In this paper we introduced a data utility-driven benchmark for selection of de-identification methods. This benchmark implements an exhaustive exploration of all combinations of de-identification methods that satisfy two key factors: adherence to the privacy requirements and minimization of data utility losses.

The benchmark provides direct support to practitioners and developers for selecting de-identification methods and making de-identification-related decisions. It also sheds light on the usage of de-identification methods and contributes to automation of de-identification processes.

Altogether, our benchmark enables better opportunities for businesses to cope with privacy-related challenges that are originating from the nature of IoT and big data.

In future steps, we will strengthen the approach with a more exhaustive overview of applicability between methods and metrics which also may consider related trade-offs and the specific nature of the data (non-structured, free text, etc.). In addition, further extension of the knowledge base, assessment of performance, investigation of opportunities for performance improvements, and conducting further evaluation of the benchmark through field tests will also be of benefit.

Acknowledgements. This research is funded by Philips Research, Research Fund KU Leuven, and the HEART project (www.heart-itn.eu). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 766139. This publication reflects only the authors' view and the REA is not responsible for any use that may be made of the information it contains.

References

1. ISO/IEC 29100 Information technology - Security techniques - Privacy framework (2011)
2. ISO 25237 - Health informatics - Pseudonymization (2017)
3. ISO/IEC 20889 - Privacy enhancing data de-identification terminology and classification of techniques (2018)
4. Abdou Hussien, A., Ramadan, N., Hefny, H.A.: Utility-based anonymization using generalization boundaries to protect sensitive attributes. J. Inf. Secur. **6**(03), 179–196 (2015). https://doi.org/10.4236/jis.2015.63019
5. Acharya, M.S.: Graduate admissions. https://www.kaggle.com/mohansacharya/graduate-admissions. Accessed 18 Apr 2019
6. Article 29 Data Protection Working Party: 0829/14/EN WP216 opinion 05/2014 on anonymisation techniques (2014)
7. Garfinkel, S.L.: NIST IR 8053: De-identification of personal information (2015)
8. Goldberger, J., Tassa, T.: Efficient anonymizations with enhanced utility. Trans. Data Priv. **3**, 149–175 (2010)
9. Ji, S., Mittal, P., Beyah, R.: Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: a survey. IEEE Commun. Surv. Tutor. **2**, 1305–1326 (2016)

10. Cao, J., Karras, P.: Publishing microdata with a robust privacy guarantee. Proc. VLDB Endow. **5**, 1388–1399 (2012)
11. Brickell, J., Shmatikov, V.: The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008)
12. Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., Radke, R.J.: A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. arXiv preprint arXiv:1605.09653 (2016)
13. Limniotis, K., Hansen, M.: Recommendations on shaping technology according to GDPR provisions - an overview on data pseudonymisation (2018)
14. LeFevre, K., De Witt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity, p. 25 (2006)
15. Sweeney, L., Samarati, P.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression (1999)
16. Sion, L., Yskout, K., Van Landuyt, D., Joosen, W.: Risk-based design security analysis. In: 2018 ACM/IEEE 1st International Workshop on Security Awareness from Design to Deployment (2018)
17. Lee, Y.J., Lee, K.H.: What are the optimum quasi-identifiers to re-identify medical records? In: 2018 20th International Conference on Advanced Communication Technology (ICACT), pp. 1025–1033. IEEE (2018)
18. Liu, Z., Qamar, N., Qian, J.: A quantitative analysis of the performance and scalability of de-identification tools for medical data. In: Gibbons, J., MacCaull, W. (eds.) FHIES 2013. LNCS, vol. 8315, pp. 274–289. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-53956-5_18
19. Machanavajjhala, A., Gehrke, J., Kifer, D.: L-diversity: privacy beyond k-anonymity. In: 22nd International Conference on Data Engineering (ICDE 2006). IEEE (2006)
20. Templ, M., Meindl, B.: Robust statistics meets SDC: new disclosure risk measures for continuous microdata masking. In: Domingo-Ferrer, J., Saygın, Y. (eds.) PSD 2008. LNCS, vol. 5262, pp. 177–189. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87471-3_15
21. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (2007)
22. Salari, M., Jalili, S., Mortazavi, R.: A utility preserving data-oriented anonymization method based on data ordering. In: 7th International Symposium on Telecommunications. IEEE (2014)
23. Nelson, G.S.: Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification. In: SAS Global Forum Proceedings (2015)
24. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity (2007)
25. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management, v0.34 (2010). http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf
26. Xiong, P., Zhu, T.: An anonymization method based on tradeoff between utility and privacy for data publishing. In: International Conference on Management of e-Commerce and e-Government. IEEE (2012)
27. Podgursky, B.: Practical k-anonymity on large datasets. Master's thesis. Vanderbilt University, Nashville, Tennessee, May 2011

28. Porter, C.C.: De-identified data and third party data mining: the risk of reidentification of personal information. 5 Shidler J.L. Com. and Tech. 3 (2008)
29. Tang, Q., Wu, Y., Liao, S.: Utility-based k-anonymization. In: 6th International Conference on Networked Computing and Advanced Information Management. IEEE (2010)
30. Ribaric, S., Ariyaeeinia, A., Pavesic, N.: De-identification for privacy protection in multimedia content: a survey. Signal Process. Image Commun. **47**, 131–151 (2016)
31. Solove, D.J.: A taxonomy of privacy. Univ. PA Law Rev. **154**(3), 477 (2006)
32. Morton, S., Mahoui, M., Gibson, P.J., Yechuri, S.: An enhanced utility-driven data anonymization method. Trans. Data Priv. **5**, 469–503 (2012)
33. The European Parliament and the Council of the European Union: Regulation (EU) 2016/679 General Data Protection Regulation. Official Journal of the European Union, pp. 1–88, May 2016
34. UC Berkeley School of Information: Privacy patterns (2018). https://privacypatterns.org/
35. Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E.W., Kantarcioglu, M., Ganta, R., Heatherly, R., Malin, B.A.: A game theoretic framework for analyzing re-identification risk. PLoS ONE **10**(3), e0120592 (2015)
36. Zuccona, G., Kotzur, D., Nguyen, A., Bergheim, A.: De-identification of health records using Anonym: effectiveness and robustness across datasets. Artif. Intell. Med. **61**, 145–151 (2014)