HySAAD—A Hybrid Selection Approach for Anonymization by Design in the Automotive Domain

Andrea Fieschi
GSaME, University of Stuttgart &
Mercedes-Benz AG
Stuttgart, Germany
andrea.fieschi@mercedes-benz.com

Pascal Hirmer, Sachin Agrawal
Mercedes-Benz AG
Stuttgart, Germany
{firstname.lastname}@mercedes-benz.com

Christoph Stach, Bernhard Mitschang
Institute of Parallel and Distributed Systems
University of Stuttgart
Stuttgart, Germany
{firstname.lastname}@ipvs.uni-stuttgart.de

***Abstract*—The increasing connectivity and data exchange between vehicles and the cloud have led to growing privacy concerns. To keep on gaining product insights through data collection while guaranteeing privacy protection, an anonymization-by-design approach should be used. A rising number of anonymization methods, not limited to the automotive domain, can be found in the literature and practice. The developers need support to select the suitable anonymization technique. To this end, we make the following two contributions: 1) We apply our knowledge from the automotive domain to outline the usage of qualitative metrics for anonymization techniques assessment; 2) We introduce *HySAAD*, a hybrid selection approach for anonymization by design that leverages this groundwork by recommending appropriate anonymization techniques for each mobile data analytics use case based on both, qualitative (i.e., "soft") metrics and quantitative (i.e., "hard") metrics. Using a real-world use case from the automotive, we demonstrate the applicability and effectiveness of HySAAD.**

***Index Terms*—Anonymization, Connected Vehicles, Privacy Protection, Metrics**

1. Introduction

In a world where data has become extremely valuable, we are in a situation where collecting data has become extremely important and the need to protect user's privacy has grown with direct proportionality.

Amongst the many possible ways of protecting privacy, there is anonymization, and in this work, we focus on *anonymization by design* as depicted in Figure 1.

A difficult step of anonymization by design is selecting the right anonymization approach. There is a multitude of very different possibilities to choose from. However, a pattern can be noticed, and approaches of different natures can at least be grouped under different clusters, e.g., grouping based or differential privacy.

This large variety of anonymization approaches can make it quite difficult to properly assess the different options, complicating the matter of having a well-justified selection during the anonymization by design process.

The lack of a technical definition of anonymization does not allow us to measure which level of anonymity has been reached. That is even more difficult to do across all techniques present in the different sets of anonymization approaches. Therefore, this is not an evaluation option during the anonymization by design cycle.

In this paper, we introduce *HySAAD* (Hybrid Selection Approach for Anonymization by Design), an approach for selecting anonymization techniques using soft and hard metrics in a top-down flow.

We define "*hard*" metrics as quantitative metrics defined to measure an aspect of an anonymization approach or simply as one that can be found used in the literature for a specific technique. It is a specific measurable aspect that cannot be shared with anonymization approaches of a different nature.

Furthermore, we introduce "*soft*" metrics to overcome the problem of not having a quantitative metric that can be applied across all the approaches that we want to evaluate as possible solutions for our data collection use case under development. We define soft metrics as evaluators for *qualitative aspects* of the anonymization approaches.

As shown in Figure 1, *HySAAD* therefore represents the first important step towards an automatic selection of a suitable anonymization technique for any analysis task. Given the requirements of a particular task as input, *HySAAD* facilitates the assessment of anonymization techniques which are hosted in a repository called anonymization toolbox.

The experience that allowed us to formulate our selection approach as well as our soft metrics comes from the automotive domain, working with experts on connected vehicles and software-defined cars. The usage of our top-down hybrid approach is however not limited to this field.

In Section 2, we outline the concept of anonymization by design as we mean it in this work. In Section 3, we show what can be found in the literature and the literature gaps that serve as motivation. We then present in Section 4 the *hard* and *soft* metrics. We explain the concept of soft metrics and list our selection that came from the expertise acquired in the automotive domain. Section 5 explains *HySAAD* and shows an example of its application using an automotive

This work is based on the research project SofDCar (19S21002), funded by the German Federal Ministry for Economic Affairs and Climate Action.

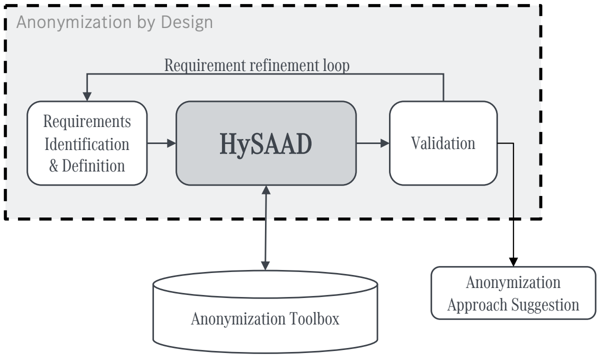
Anonymization by Design
Requirement refinement loop
Requirements Identification & Definition
HySAAD
Validation
Anonymization Toolbox
Anonymization Approach Suggestion

Figure 1: HySAAD selection method placed in an anonymization by design cycle

domain use case. Finally, Section 6 concludes this paper and outlines future research.

2. Anonymization By Design

An effective way of protecting privacy is to do that by design, e.g., following the privacy by design principles [1]. Privacy protection should be considered since the early steps of a design process, not only in hindsight as a Privacy Enhancing Technology (PET) [2]. The same goes for anonymization. In this work, we talk about anonymization by design. Anonymization by design can be structured as a cycle. In Figure 1, we show the anonymization by design cycle and where *HySAAD* fits. The cycle is iterative. It starts from identifying the requirements and providing those to a selection mechanism. This selection mechanism has access to a toolbox of anonymization approaches and returns a recommendation based on the requirements received. The recommended anonymization approach is then validated. This step takes into consideration if the approach is feasible, if it reflects the original idea, or if the recommendations received are still too imprecise. With the output of the validation step, we loop back to the first step, refine the requirements, and repeat the cycle. The mobile environment and its many service-oriented applications need to protect users' privacy. Anonymization by design can help developers provide stronger privacy protection.

3. Related Work

The presence of the topic of anonymization is substantial in the literature. The pattern that emerges is that there are clusters of anonymization approaches that use the same idea to achieve a property that can be used to guarantee anonymity. In this paper, we consider two sets of anonymization approaches: Grouping-based approaches, and differentially private approaches. The former set of approaches guarantees anonymity through the creation of a group of data sources that are indistinguishable from one another, e.g., k-anonymity [3]. The latter set comprehends approaches that aim at satisfying Differential Privacy (DP) [4].

We searched in the literature for lists of anonymization metrics and comparisons of different data anonymization approaches. This was aimed at giving us an understanding of the metrics used for anonymization evaluation.

In the literature, we found works that provide a survey of privacy concepts and PETs for the cloud environment [5] and for tabular and graph data [6]. We also encounter works that present multiple anonymization approaches for k-anonymity and measure their efficiency concerning resource usage and effectiveness about data utility [7].

We found works trying to relate t-closeness and DP either extending the former and generalizing the latter [8], [9], or proving that t-closeness implies DP and DP implies a stochastic version of t-closeness.

The use of qualitative metrics, in this work referred to as *soft metrics*, is also present in the literature. Evaluating qualitative features has been proven to be helpful to evaluate environments for visual computing [10] or benchmarking specialized hardware for deep learning algorithms [11], [12].

We found a literature gap that leaves uncovered the possibility of comparing anonymization approaches of different natures, such as approaches of a differentially private nature and all the others of the grouping-based cluster, or data synthesis and compound approaches using encryption.

4. Anonymization Metrics

In a broad sense, a metric is a measure used to assess or evaluate a particular characteristic, performance, or quality of a system, process, product, or service [13]. Metrics can even be used to motivate choices made during the design process and show how an evaluation using said metrics leads the designing group towards one direction instead of the other. Every field has its own metrics to quantify and analyse different aspects of performance or effectiveness. The word metric is usually associated with a quantifiable measurement, something that can quantitatively help us to evaluate our characteristics under analysis. In this work, we refer to quantitative metrics as "*Hard* Metrics". In Section 4.1, we show a set of hard metrics for anonymization that practical experience has shown to be the most relevant. However, excluding the element of "quantifiable measurability" from the description of metrics, we were able to expand our horizon and look at qualitative metrics too, which we decided to name "*Soft* metrics" in the context of *HySAAD* and this work. In Section 4.2 we describe to the reader the soft metrics that were selected for *HySAAD*.

4.1. Hard Metrics, Quantitative Analysis

With the aid of the PRISMA method [14], we collected several anonymization approaches. As already explained in Section 3, the metrics found in the literature are specific for a certain kind of anonymization approach and rarely can work across different anonymization approaches sets like DP and grouping-based. Table 1 gives an overview of the eleven metrics presented in this section and their characteristics. The characteristics highlighted in the table are further explained in the following paragraphs.

The "Type" of a metric has one of four values. *Syntactic* metrics measure the structural properties of the anonymized dataset. These metrics do not consider the attribute values, they evaluate syntactic properties, not semantic ones, e.g., the sizes of Equivalence Classes (ECs). Metrics of this type can usually be calculated with low computational costs. *Distance-based* metrics measure a notion of distance between records or cell values. *Distribution-based* metrics are semantic measures, they consider the values and distributions of these values in the table. *Adversarial* metrics consider the resilience to an adversarial attack.

The property "Measure" indicates what aspect of the anonymized data is considered by the metric. Metrics that evaluate the performance of any given anonymization approach in the literature mainly examine data utility (U) and privacy (P). Data utility relates to the ability to learn aggregate statistics about large groups of individuals. Privacy relates to the ability to extract information about specific individuals in the data [24], [25].

The "Granularity" indicates the level of detail with which the individual entries are considered in the table. Metrics that operate on a cell level consider the individual attribute values, while record-based metrics operate on a tuple-by-tuple basis.

"Relative" indicates whether the metric returns a score relative to a baseline dataset. Metrics that measure utility can incorporate the utility of the raw data in their calculation to capture information loss due to anonymization. Metrics that measure privacy can incorporate the privacy of the trivially anonymized dataset, where all quasi-identifier values are removed [24]. Relative metrics can convey more information. For example, if the utility of the raw data is low, it will consequently also be low in the anonymized data. A relative metric takes into consideration the data utility of the original data set and highlights the loss caused by the applied anonymization approach.

Table 1: Overview of the selected metrics and their characteristics.

	Type	Measure	Granularity	Relative
Suppression Ratio [15]	Syntactic	P + U	Record	Y
Minimum k [16]	Syntactic	P + U	Record	N
Minimum l [17]	Syntactic	P	Cell	N
Average Equivalence Class Size [17]	Syntactic	P + U	Record	N
Discernibility Penalty [3]	Syntactic	P + U	Record	N
In-Data Precision Loss [18]	Distance	P + U	Cell	N
Cross-Data Precision Loss [19], [20]	Distance	P + U	Cell	Y
Earth Mover's Distance [21]	Distribution	P	Cell	Y
g-balance [22]	Distribution	P	Record	N
h-affiliation [22]	Distribution	P	Record	N
Adversarial Knowledge Gain [23]	Adversarial	P	Cell	Y

In Table 2, we can see on which anonymization approaches we can apply the selected hard metrics. We list the applicability for four different approaches: 1) k-anonymity through generalisation, 2) k-anonymity through microaggregation, 3) data synthesis, and 4) differential privacy.

To indicate the full applicability of the metric with the anonymization approach, we use a "yes". If the metric does not work with the anonymization approach we use a "no". To highlight where the metric can be applied to the method but loses expressiveness, we use a "$\sim$".

Some interesting things can be easily extrapolated from Table 2. The first one is that all the selected metrics can be used when Data Synthesis is the anonymization approach of choice, however all but Cross-Data Precision Loss lose expressiveness when applied to it, making it difficult to claim that it could be used to properly evaluate k-anonymity and Data Synthesis for example. Then we, also notice that none of the selected metrics can be used when DP is the anonymization approach of choice. This was already expected since in Section 3, we already exposed the lack of generally applicable metrics as a literature gap.

Table 2: Applicability of metrics on methods.

	k-anonymity through Generalization	k-anonymity through Microaggregation	Data Synthesis	Differential Privacy
Suppression Ratio [15]	yes	yes	$\sim$	no
Minimum k [16]	yes	yes	$\sim$	no
Minimum l [17]	yes	yes	$\sim$	no
Average Equivalence Class Size [17]	yes	yes	$\sim$	no
Discernibility Penalty [3]	yes	yes	$\sim$	no
In-Data Precision Loss [18]	yes	$\sim$	$\sim$	no
Cross-Data Precision Loss [19], [20]	yes	yes	yes	no
Earth Mover's Distance [21]	yes	yes	$\sim$	no
g-balance [22]	yes	yes	$\sim$	no
h-affiliation [22]	yes	yes	$\sim$	no
Adversarial Knowledge Gain [23]	yes	yes	$\sim$	no

Note: "yes": applicable; "no": not applicable; "$\sim$": partial applicability.

As a hard metric for DP, we could use its definition. The definition of DP [4] can be used as a metric on its own not only as a goal to be achieved by an algorithm. The privacy parameter ϵ is also referred to as the *privacy budget* in the literature [26].

Therefore, we can see that on top of having metrics that were explicitly thought for measuring aspects of specific anonymization approaches, we also have their parameters that can give us a direct indication of how much the approach will affect the data. Same as for the approach-specific metrics, these parameters will not allow us to compare approaches of different natures, but it will allow us to understand how much protection (or distortion) a specific approach will provide.

If we think about DP, the parameter ϵ directly gives us an idea of how much the data will be distorted and how difficult it will be for private information to be disclosed. The lower the privacy budget, the higher the privacy protection as well as added distortion and information loss.

We can consider the k in k-anonymity, which in the same way tells us how much the data has been modified and how difficult it will be to identify the source of a data entry.

Both parameters k and ϵ have this inherent measurability and quantifiability that can be used as a measure for comparisons between methods of the same family but with different parameter settings or slightly different approaches, hence, they can be used as hard metrics.

4.2. Soft Metrics, Qualitative Analysis

One of our suggestions with *HySAAD* is to remove the aspect of "quantifiable measure" from the definition of metric.

In this work, *soft metrics* are qualitative aspects that can be used as metrics. Those by nature are general and applicable to any kind of anonymization approach coming from any class of anonymization.

When it comes to privacy protection, there is a generally undervalued characteristic which is the *gut feeling* perception. As users, we can all recognize how intuition leads us to be more careful about our privacy in one situation and not in the other. The general understanding of the privacy-violation-potential helps people be more careful.

Let us think about how we choose our passwords and how we protect our profiles on different apps, websites, and portals: when it comes to home banking most people would be more careful compared to when they set up a profile for an online poker website. The overhead that privacy protection might cause also pushes harder the risk of users caring about their privacy only when the effort seems to be worth it [27]. The same intuition or gut feeling helps in the evaluation of privacy-protecting techniques or privacy enhancing techniques (PETs), too.

Anonymization already generally "feels" more privacy-protecting; as opposed to some PETs that require the user to trust that, for example, the server where their data are saved are monitored and protected. With anonymization, once the data has been guaranteed to be anonymous, the user can stop worrying about people finding something about them. Now, that guarantee of anonymity is the problem. Stripping the data of personally identifiable information (PII) is not enough, not only from a technical point of view but also to convince the user of their anonymity.

Let us imagine the typical anonymization technique that is used for video interviews. The first step would be blurring the face of the interviewee. However, this might not feel like it is enough: the voice could still be recognized by people that already heard the interviewee speak. Hence, the voice can be altered to enhance the feeling of effective anonymity. If the topic of the interview is particularly sensitive, the subject in the video might also argue that the surrounding might give away their identity; therefore, a dark studio with no identifying features would be preferable. Also, clothing and inserts that will alter the identifiability of physical features might be employed. Not it becomes even more difficult to infer the presence of identifying body features or the body type of the subject, e.g., tall or short, full-figured or lean.

This example helps us understand how certain approaches can be deemed as "enough" in certain cases and "too weak" in others, and how this can only be qualitatively evaluated.

To effectively consider different anonymization approaches, we also need to consider in which situation they are applied, which fields, and which kind of feeling of protection they provide. The feeling of protection can sometimes be helped by the *clear understandability* of an approach, e.g., it is effective and clear to understand, or it could be helped by pure semantics and nomenclature, e.g., they simply sound effective and secure.

4.2.1. Selected Soft Metrics.

After interviews with experts and with the experience gained in the automotive domain, we have derived the following soft metrics:

SM1: Raw Data Permanence
SM2: Required Amount of Data Preprocessing
SM3: Data Use Case Rigidity
SM4: Intuitiveness of Privacy Effects
SM5: Required Amount of Domain Knowledge
SM6: Unsuitability for on-line Data Collection
SM7: Ease of Application
SM8: Complexity of Implementation

In the following, we give a more extensive explanation of what the aforementioned soft metrics mean.

SM1-Raw Data Permanence. Some techniques for privacy protection require permanent access to the raw personalized data, which then needs to be stored securely in a databank which can guarantee those security standards. For example, techniques for Privacy Preserving Data Analysis tend to require permanent access to the raw data.

SM2-Required Amount of Data Preprocessing. If an anonymization approach has strict requirements regarding the structure of data it expects, preprocessing the raw data may be a substantial task. The transformation of the raw data may introduce human error or result in information loss.

SM3-Data Use Case Rigidity. Approaches for Privacy Preserving Data Publishing (PPDP) release an anonymized dataset to allow a wide range of data use cases. On the

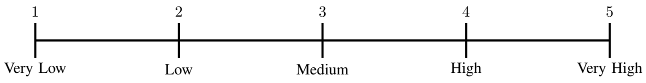

Figure 2: Five-point scale for soft metrics

other hand, techniques for Privacy Preserving Data Analysis (PPDA) target very specific data use cases. This makes the usability of their outputs narrow but allows tailoring the mechanism exactly to a given use case. The selection of a fitting anonymization approach may be influenced by assessing whether the data will have broad or specific uses.

SM4-Intuitiveness of Privacy Effect. Effective communication regarding the privacy measures taken to protect customer data greatly benefits companies. Customers who understand how their private data is managed build trust in the company. On the other hand, effective communication allows fast resolution of any inquiries coming from legal institutions. This metric captures how intuitive the privacy mechanism is to a non-specialist. For later purposes, we attribute this soft metric the following name: $I.P.E.$

SM5-Required Amount of Domain Knowledge. This soft metric assesses the difficulty level involved in implementing one of the anonymization approaches by someone unfamiliar with the data domain. A low requirement indicates that a data engineer can apply an anonymization approach to different types of data without the need to invest time to understand the semantics and structure of the data.

SM6-Unsuitability for on-line Data Collection. Often, new data arrive continuously, and anonymization approaches should be able to handle the steady stream of new data. This metric assesses the ability of an anonymization approach to anonymize new data efficiently and effectively.

SM7-Complexity of Implementation. Some anonymization approaches can be extremely challenging to implement, also when the idea is relatively intuitive, the actual implementation and required computational power can increase the complexity of the data collection.

The soft metrics listed above are broad enough to apply across various anonymization methodologies, irrespective of their specific categorization. Their qualitative assessment does not deal with the technical details of each approach, instead, it looks at factors that cover how effective and applicable the anonymization process is. To give a qualitative rating for the aspects evaluated by the soft metrics, we can use the five-point scale system shown in Figure 2. The experience in the automotive domain and exchanges with experts led us to settle for five points instead of more or fewer. Being a qualitative rating given with feeling and not with a measure, a five-point scale seems to be the best option, not too finely grained but still allows the developer to express its judgement, in the same way that it is done for Quality Function Deployment (QFD) analysis [28].

Soft metrics can be redefined and extended. Depending on the field of application, specific needs, or simply the further experience acquired. They can be tailored to the specific field of application, but they are a good starting point for any other situation.

It is interesting to notice how almost all seven soft metrics have a “negative” element connected to them, or to be more precise, using the five-point scale, a 5 would generally be bad or more complicated while a 1 would generally indicate a better or easier anonymization approach. That is however not true for “Intuitiveness of Privacy Effects”. This is the only soft metric that moves the other way. We decided, anyway, not to change it in order to have a more meaningful name. We can simply use the complement $(I.P.E.)'$ of the soft metric, which becomes:

$$(I.P.E.)' = 6 - I.P.E.$$

5. HySAAD: Hybrid Selection Approach for Anonymization by Design Cycles

The soft metrics described in Section 4.2 can be applied to all anonymization approaches regardless of how they can be clustered. Their general and non-quantitative nature allows the developers to make a first choice between the different classes of anonymization approaches, i.e., differential privacy, grouping based, etc. Once this first choice has been made, the details of the approach and its settings can be evaluated using metrics that work well for the selected class of anonymization approaches. Soft metrics are not limited to the ones listed in Section 4.2, new ones can be formulated with more experience and with domain-specific interests in mind. Their use would remain the same: giving a first skimming opportunity amongst the many possible anonymization approaches and then helping to explain the design choices taken during development. Soft metrics are more understandable and convey more information to people that are not familiar with the field, or to end customers.

5.1. HySAAD Overview

We can see in Figure 3 how the *HySAAD* is structured. To understand how *HySAAD* fits in the anonymization by design cycle we take a look at Figure 1. Our selection approach takes the requirements, has access to an anonymization toolbox, and through the iterative use of soft and hard metrics, it returns anonymization approach suggestions. The reader should keep in mind that, especially during the first iterations of the anonymization by design cycle, the suggestion will not be unique. With the anonymization approaches given by *HySAAD*, the requirements are refined and a new iteration of the anonymization by design cycle starts. The metrics, in particular the soft ones, give the developer some sort of an explanation and “weight” of why those approaches were selected as possible solutions. *HySAAD* is a selection approach that, as the name says, is hybrid and uses iteratively soft and hard metrics, but on top of its iterative nature, it also takes into account that it will be used in an iterative environment, namely the anonymization by design cycle. When no solution is strongly better than another *HySAAD* can easily suggest multiple solutions, with

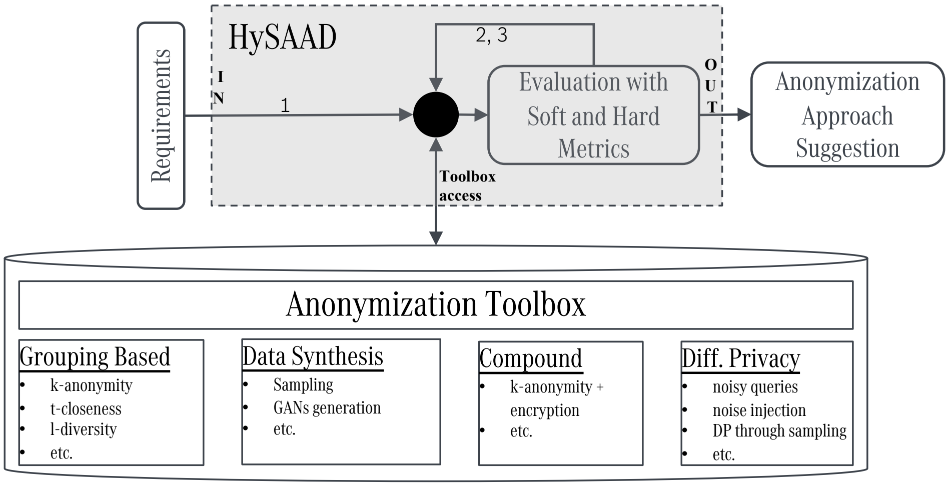

Figure 3: HySAAD selection method structure. The first, second, and third iterations of the example in Section 5.2.3 are marked with numbers.

its reasons attached, i.e., the soft metrics, and the designer can use those to refine the requirements. Once the requirements become more precise and they steer the design process to a specific direction then out of the *HySAAD* the developer can extrapolate a specific suggestion, with the hard metrics helping to define the value of the parameters.

This process can be aided by the QFD method [28], [29]. The QFD method can help the designer during the evaluation process in the *HySAAD* loop and in the anonymization by design cycle too.

5.2. HySAAD Evaluation on Industry Use Case

For the purpose of evaluation, we apply the *HySAAD* selection approach to an example use case coming from the automotive domain. We can there then explain how we would apply our *soft metrics* in this case. First, we draft an example anonymization toolbox and an example use case.

5.2.1. Toolbox. Let us consider as available for this example the methods shown in Table 3 and let that be our toolbox for our anonymization by design cycle, hence, the toolbox that *HySAAD* will have access to in this example.

Table 3: Toolbox

Cluster	Approach	Specific Implementation
Grouping-Based Anonymity	k-Anonymity [30]	Generalisation [31]
		Suppression [15]
	l-Diversity [32]	As proposed in [17]
Differential Privacy [33]	Noise Injection	Differentially Private Queries [34]
		Randomized Response Mechanism [35]
Data Systhesis [36]	New dataset with same distribution as the original	GANs [37]
		Sampling [38]

5.2.2. Use Case. To investigate the Human Computer Interaction (HCI) issues, an HCI expert is interested in collecting the data on how many times the touchscreen is used during a given drive. This would provide important information of how much the system gets used and how useful it is deemed by drivers. We collect data to know how the product works and is used across the whole fleet, and we do not need user identifiability [39]. This tells us that anonymization is a possible privacy protection approach. We use an anonymization by design approach and we follow the scheme shown in Figure 1.

5.2.3. Using HySAAD. Knowing our goal, i.e., the use case above explained and anonymization as privacy protection, *HySAAD* first uses *soft metrics* to decide amongst the different classes of anonymization methods and then it passes to *hard metrics* in later iterations. All the soft metrics presented in Section 4.2 are considered. However, in this case, we decide to give particular importance to 1) Raw Data Permanence with a low score to prevent the permanence of non-anonymized data in the cloud. 2) $(I.P.E.)'$ with a low score so that a simple explanation can be understandable by management and users. 3) Complexity of Implementation with a low score to ensure the design of a use case that can be quickly implemented and tested.

First Iteration. Let us start with evaluating the anonymization approaches available in the toolbox. In Table 4, we see how we can use the soft metrics during the first iteration of *HySAAD* for a first assessment of the anonymization methods available in the toolbox. We first compare the different sets of anonymization approaches present in our toolbox. For each soft metric, we use the five-point scale from Figure 2 to give a qualitative rating.

Table 4: First iteration of HySAAD, soft metrics over the families of anonymization approaches given in the toolbox

	Grouping Based	Differential Privacy	Data Synthesis
Raw Data Permanence ⋆	2	5	3
Required Amount of Data Preprocessing	2	1	3
Data Use Case Rigidity	3	4	2
$(I.P.E.)'$ ⋆	1	4	3
Required Amount of Domain Knowledge	3	2	2
Unsuitability for on-line Data Collection	3	1	4
Complexity of Implementation ⋆	1	4	3
Average	2.14	3	2.86
Average of ⋆ metrics	1.33	4	3

The grouping-based cluster of anonymization approaches has the lowest average score, which can already be an indication that this type of anonymization approach will be less problematic in this situation. We can also highlight specific elements that seem more important, quite like attributing different weights to the soft metrics. We therefore opt for grouping-based anonymization approaches since this set yields an overall lower score and also for the soft metrics that we highlighted in this example the scores are lower.

Second Iteration. Now, we can do a second iteration and consider the specific approaches that we have in the Grouping-Based set. Also, during this second iteration, we use the soft metrics. We can see the scores attributed during the second iteration in Table 5.

At the end of this second iteration considering the average score, and the single scores of the soft metrics, we decided to give more importance to, we can see that k-anonymity seems to be the best option.

Third Iteration. We switch now to hard metrics and our choice is between k-anonymity through generalization and k-anonymity through microaggregation. Considering our data set and the use case we have, and as an illustrative example, we decided to set $k = 3$. We pick the following as indicative hard metrics: 1) Suppression Ratio 2) Minimum k 3) Average Equivalence Class Size.

Table 5: Second iteration of HySAAD, soft metrics over the grouping-based approaches

	k-Anonymity	l-Diversity
Raw Data Permanence ⋆	2	2
Required Amount of Data Preprocessing	2	3
Data Use Case Rigidity	3	3
$(I.P.E.)'$ ⋆	1	2
Required Amount of Domain Knowledge	3	3
Unsuitability for on-line Data Collection	3	2
Complexity of Implementation ⋆	1	2
Average	2.14	2.43
Average of ⋆ metrics	1.33	2

Table 6: Third iteration of HySAAD, hard metrics over the specific implementations

	k-anonymity through Generalization	k-anonymity through Microaggregation
Suppression Ratio	$2/10$	$3/10$
Minimum k	3	3
Average Equivalence Class Size	4	3

As we can see in Table 6, k-anonymity through generalization has a lower Suppression Ratio and a higher Average Equivalence Class Size. This makes it the better candidate for a possible first suggestion.

HySAAD Output. *HySAAD* would at this point output k-anonymity through generalization with $k = 3$ as a suggestion. The anonymization by design cycle would take this suggestion, validate it, and go back to the stage of refinement of the requirements. The refined requirements are then given as input to *HySAAD* and the process repeats.

Observations. *HySAAD* allows anonymization by design to be effectively used in the automotive field when anonymized data want to be collected and later managed. With its combination of soft and hard metrics, it is of great support to developers and allows the anonymization by design cycle to be successfully used.

6. Conclusion & Outlook

In this work, we have seen how, even with the lack of quantitative metrics, it is possible to compare anonymization techniques of different natures using the soft metrics we presented here in order to analyze qualitative features. Our soft metrics also allowed us to conceive *HySAAD*, make it possible to assess anonymization techniques, and select the most appropriate one for a given task and thereby realize the anonymization by design cycle. Mobile data collection, in the automotive domain or in other fields, can improve its reach and users' privacy protection through anonymization by design. *HySAAD* is the formalization of a central step of the anonymization by design cycle.

In future work, we aim to automate our method to increase its applicability in real-world scenarios. This also entails the automated calculation of hard metrics for newly added anonymization techniques. Moreover, we aim to apply our privacy by design method to real-world application scenarios.

References

[1] M. Langheinrich, "Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems," in *Proceedings of the 2001 International Conference on Ubiquitous Computing (Ubicomp)*, 2001.

[2] A. Morton and M. A. Sasse, "Privacy is a process, not a PET: A theory for effective privacy practice," in *Proceedings of the 2012 New Security Paradigms Workshop (NDPW)*, 2012.

[3] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, 2005.

[4] C. Dwork, "Differential Privacy: A Survey of Results," in *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC)*, 2008.

[5] P. Silva, E. Monteiro, and P. Simões, "Privacy in the Cloud: A Survey of Existing Solutions and Research Challenges," *IEEE Access*, vol. 9, pp. 10 473–10 497, Jan. 2021.

[6] A. Majeed and S. Lee, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," *IEEE Access*, vol. 9, pp. 8512–8545, Dec. 2020.

[7] V. Ayala-Rivera *et al.*, "A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners," *Transactions on Data Privacy*, vol. 7, no. 3, pp. 337–370, Dec. 2014.

[8] J. Domingo-Ferrer and J. Soria-Comas, "From t-closeness to differential privacy and vice versa in data anonymization," *Knowledge-Based Systems*, vol. 74, pp. 151–158, Jan. 2015.

[9] E. Ekenstedt *et al.*, "When Differential Privacy Implies Syntactic Privacy," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2110–2124, May 2022.

[10] J. Scholtz, "Developing qualitative metrics for visual analytic environments," in *Proceedings of the 3rd BELIV'10 Workshop: BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization (BELIV)*, 2010.

[11] K. Ovtcharov *et al.*, "Accelerating Deep Convolutional Neural Networks Using Specialized Hardware," Microsoft Research, White Paper, Feb. 2015. [Online]. Available: https://www.microsoft.com/en-us/research/publication/accelerating-deep-convolutional-neural-networks-using-specialized-hardware/.

[12] W. Dai and D. Berleant, "Benchmarking Contemporary Deep Learning Hardware and Frameworks: A Survey of Qualitative Metrics," in *Proceedings of the 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, 2019.

[13] F. Rahman and P. Devanbu, "How, and why, process metrics are better," in *Proceedings of the 2013 35th International Conference on Software Engineering (ICSE)*, 2013.

[14] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *The BMJ*, vol. 372, pp. 1–9, 2021.

[15] P. Samarati and L. Sweeney, "Protecting Privacy when Disclosing Information: K-Anonymity and its Enforcement through Generalization and Suppression," Computer Science Laboratory, SRI International, Technical Report SRI-CSL-98-04, 1998. [Online]. Available: http://www.csl.sri.com/papers/sritr-98-04/.

[16] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[17] A. Machanavajjhala *et al.*, "L-diversity: Privacy beyond k-anonymity," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, 2006.

[18] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

[19] J. Soria-Comas *et al.*, "t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3098–3110, May 2015.

[20] H. Kikuchi *et al.*, "Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization," in *Proceedings of the 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, 2016.

[21] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering (ICDE)*, 2007.

[22] H. B. Kartal and X.-B. Li, "Protecting Privacy When Sharing and Releasing Data with Multiple Records per Person," *Journal of the Association for Information Systems*, vol. 21, no. 6, pp. 1461–1485, 2020.

[23] J. Brickell and V. Shmatikov, "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 (KDD)*, 2008.

[24] 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 (KDD)*, 2009.

[25] G. Cormode *et al.*, "Empirical privacy and empirical utility of anonymized data," in *Proceedings of the 2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 2013.

[26] T. Zhu *et al.*, "Differentially Private Data Publishing and Analysis: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1619–1638, Aug. 2017.

[27] M. Ramokapane, A. C. Mazeli, and A. Rashid, "Skip, Skip, Skip, Accept!!! A Study on the Usability of Smartphone Manufacturer Provided Default Features and User Privacy," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 2, pp. 209–227, 2019.

[28] L.-K. Chan and M.-L. Wu, "A systematic approach to quality function deployment with a full illustrative example," *Omega*, vol. 33, no. 2, pp. 119–139, Apr. 2005.

[29] L.-K. Chan and M.-L. Wu, "Quality Function Deployment: A Comprehensive Review of Its Concepts and Methods," *Quality Engineering*, vol. 15, no. 1, pp. 23–35, 2002.

[30] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571–588, Oct. 2002.

[31] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, Aug. 2001.

[32] A. Machanavajjhala *et al.*, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, 3–es, Mar. 2007.

[33] C. Dwork, "Differential Privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, 2006.

[34] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, Aug. 2014.

[35] Y. Wang, X. Wu, and D. Hu, "Using Randomized Response for Differential Privacy Preserving Data Collection," in *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference (EDBT/ICDT Workshops)*, 2016.

[36] C. M. Bowen and F. Liu, "Comparative Study of Differentially Private Data Synthesis Methods," *Statistical Science*, vol. 35, no. 2, pp. 280–307, May 2020.

[37] N. Park *et al.*, "Data synthesis based on generative adversarial networks," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, pp. 1071–1083, Jun. 2018.

[38] B. C. M. Fung *et al.*, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 2010.

[39] A. Fieschi *et al.*, "Anonymization Use Cases for Data Transfer in the Automotive Domain," in *Proceedings of the 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 2023.