Comparison of De-Identification Techniques for Privacy Preserving Data Analysis in Vehicular Data Sharing

Sascha Löbner
Frédéric Tronnier
Sebastian Pape
Kai Rannenberg
sascha.loebner@m-chair.de
frederic.tronnier@m-chair.de
sebastian.pape@m-chair.de
kai.rannenberg@m-chair.de
Chair of Mobile Business & Multilateral Security, Goethe University Frankfurt
Frankfurt am Main, 60323

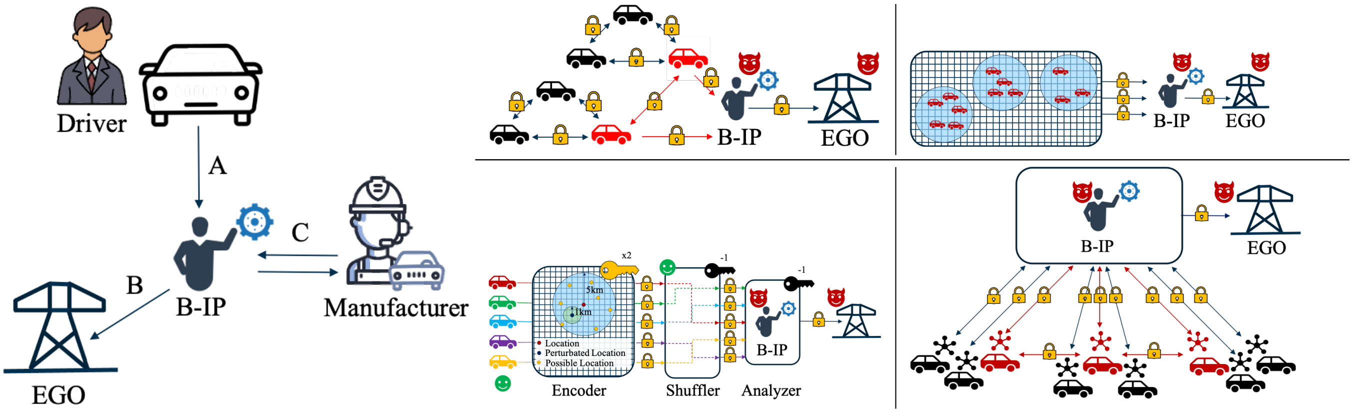

Figure 1: Data flow chart and de-identification techniques, including the Vehicle/Driver, Business Intelligence Provider (B-IP) and Energy Grid Operator (EGO).

ABSTRACT

Vehicles are becoming interconnected and autonomous while collecting, sharing and processing large amounts of personal, and private data. When developing a service that relies on such data, ensuring privacy preserving data sharing and processing is one of the main challenges. Often several entities are involved in these steps and the interested parties are manifold. To ensure data privacy, a variety of different de-identification techniques exist that all exhibit unique peculiarities to be considered. In this paper, we show at the example of a location-based service for weather prediction of an energy grid operator, how the different de-identification techniques can be evaluated. With this, we aim to provide a better understanding of state-of-the-art de-identification techniques and the pitfalls to consider by implementation. Finally, we find that the optimal technique for a specific service depends highly on the scenario specifications and requirements.

CCS CONCEPTS

• **Security and privacy** → **Data anonymization and sanitization**; **Security requirements**.

KEYWORDS

privacy, de-identification, anonymization, autonomous vehicles, automotives, privacy preserving data analysis, data sharing

ACM Reference Format:
Sascha Löbner, Frédéric Tronnier, Sebastian Pape, and Kai Rannenberg. 2021. Comparison of De-Identification Techniques for Privacy Preserving Data Analysis in Vehicular Data Sharing. In *Computer Science in Cars Symposium (CSCS '21), November 30, 2021, Ingolstadt, Germany.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3488904.3493380

1 INTRODUCTION

The ongoing digitalization of automotive vehicles is partially driven by the desire of autonomous driving. However, already in its current state, automotives form a swarm of moving sensors, permanently

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CSCS '21, November 30, 2021, Ingolstadt, Germany
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9139-9/21/11…$15.00
https://doi.org/10.1145/3488904.3493380

recording various kinds of data. Thus, an obvious idea is to make use of this data for other purposes than (autonomous) driving. For example, the data could be integrated into concepts for smart cities. A report from McKinsey already discussed in 2016 the monetization of car data [7] and recently, Inrix announced their data marketplace IQ[1] for anonymized location-based data [46].

However, anonymizing location-based data is not an easy task and can easily lead to privacy or compliance violations. If the related data can be used to identify the vehicle, the driver or the passengers it has to be considered as personally identifiable information (PII). E. g. [37] showed in the context of individual insurance models that the identification of a driver in a group of all users of a vehicle was possible with more than 90% accuracy. Due to the General Data Protection Regulation (GDPR) [16], companies need to provide documentation for the explicit consent of EU citizens if they want to collect and process PII. This is also in line with a set of privacy principles by the Alliance of Automobile Manufacturers (AAM) from 2014 [1] which encourages affirmative consent for the collection of sensitive data such as the driver's biometrics, geolocation or driver behavior data. Pesé and Shin [30] provide an overview of relevant automotive privacy regulations. A report from 2017 [2] seems to confirm that car manufactures aim to respect these guidelines.

Getting the driver's consent may not always be easy for the car manufacturer since the driver does not need to be the owner of the car. Thus, de-identification, preventing the identification of persons from the collected data, avoids the need to ask for consent. While the consumers' intention to use a car is still mostly influenced by the perceived benefits and privacy risks only have a minor influence [9], legal and compliance issues – as sketched above – provide a strong incentive to get the de-identification right.

On the other hand, even though several anonymisation techniques are available and the European Data Protection Board (EDPB) issued guidelines on the processing of personal data in the context of connected vehicles including guidance on data anonymisation [3], a public consultation on the content of these guidelines re-iterated the need for clearer guidance with good practices of data anonymisation [10]. The anonymization process rises several challenges: i) Proving that the collected data is properly de-identified, which involves proving that any re-identification is impossible. ii) The average on-the-road lifespan of a vehicle is about 11 years with roughly 5 years before needed to design the vehicle [19]. That requires foresight of almost two decades or a process for regular updates. iii) With several de-identification methods in place[2] and a magnitude of possible combinations, it is difficult to select the best or even a propper approach.

Certainly, there is no one-size-fits-all approach and the used approach needs to be aligned to the relevant scenario. In this paper we aim to identify and evaluate suitable de-identification approaches for a scenario in which weather data is collected by cars and shared with the operator of an energy grid to allow the operator more reliable forecasts for the production of renewable energy (cf. Sect. 4). The scenario is meant to be a realistic example but the lessons learned in the analysis are not limited to the specific scenario and can be transferred to related scenarios as well.

Our contribution is the presentation of an elaborated scenario description (in Sect. 4) in which collected data is shared with a third party without the need to ask the driver for consent. Furthermore, we elicit requirements for the suitable de-identification approaches based on our threat model (Sect. 5) and then discuss (dis-)advantages of the considered de-identification approaches (Sect. 7).

2 BACKGROUND AND RELATED WORK

In this section we provide a brief overview of existing de-identification techniques and related literature.

2.1 De-identification

This section aims to create a comprehensive overview on the current status of academic literature on de-identification methods and techniques. In the following, an overview on the models available in ISO/IEC 20889:2018 and the privacy preserving machine learning literature is provided.

In general, the de-identification techniques from the ISO/IEC 20889:2018 [21] are separated into 8 major classes. First, statistical tools such as sampling and aggregation. Second, cryptographic tools including deterministic, order-preserving, and Homomorphic Encryption (HE), as well as secret sharing. Third, suppression, including masking, local suppression, record suppression and sampling. Four, pseudonymization aiming to replace original identifying attributes with independent pseudonyms, e.g. using randomization. Five, granularization reducing the granulartiy of information in attributes with techniques such as rounding or top/bottom coding. Six, randomization modifying attributes randomly utilizing e.g. noise addition, permutation, or micro aggregation. Seven, differential privacy, a system to share information in a dataset while withholding information about a single information in that dataset [15]. Eight, k-anonymity defining a state where a person cannot be distinguished from $k-1$ other persons in a dataset [38].

Especially for machine learning scenarios, Al-Rubaie and Chang [4] expand that list by secure processors, also known as Trusted Execution Environments (TEE) ensuring the confidentiality and integrity of the source code. Moreover, in the domain of privacy preserving machine learning, Secure Multiparty Computation (MPC) is a well-known approach to ensuring data privacy when computing data from multiple sources [11, 25, 34]. Another technique for privacy preserving machine learning is Federated Learning (FL) that enables the training of a model on a local device without sharing all data with a central entity [25, 44].

2.2 Related Work

In the following, prior scientific research on de-anonymisation techniques are presented that are of importance to this work.

Gruteser and Grunwald [20] analyse the anonymous usage of vehicular location-based services, such as fleet management, traffic monitoring or consumption-based car insurances that are based on telematics. They conclude that the risk of re-identification and location tracking can be reduced utilizing k-anonymous data. A significant amount of research has also been conducted on mobility-related use cases such as VANETs using k-anonymity [40, 41] or HE [17, 31] to ensure the privacy of vehicles. Pape and Rannenberg [29] demonstrate how the application of privacy patterns in fog

[1] https://inrix.com/products/inrix-iq/

[2] ISO/IEC 20889:2018 [21] lists more than 20 different approaches

computing environments can improve the users' privacy in a smart vehicle use case.

Frank and Kuijper [18] investigates vehicle users' privacy concerns by evaluating the use of cameras and capacitive proximity sensing in driver assistance systems. As a result of their survey they find evidence that the anonymization by capacitive proximity sensing is preferred. Thereby, they underline the impotency to address the privacy concerns of vehicle users by sufficient technical solutions. Also Krontiris et al. [22] find that the consumers acceptance of autonomous vehicles depends on a privacy preserving design that protects against tracking.

Krumm [23] compares different de-identification techniques for inference attacks on location tracks utilizing an experimental assement. He claims that the required degree of corruption for noise or rounding is very likely to make location-based services unusable. In his test environment, spatial cloaking based on k-anonymity was only effective within a 2 km radius. Ribaric et al. [35] reviews techniques for de-identification of personal identifiers in another context: multimedia contents. They classify personal identifiers into non-biometric, biometric and soft biometric identifiers.

Kumar et al. [24] review and compare existing techniques for de-identification with the aim to protect the personal privacy. In their conclusion they line out that k-anonymity, 1-diversity and T-closeness can reduce the risk of personal data unveiling although they are vulnerable against some privacy attacks. Also Murthy et al. [28] analyze and compare perturbation, anonymization and cryptographic approaches. They conclude that from the compared techniques, suppression stands out while swapping lags behind due to massive resource consumption. Al-Rubaie and Chang [4] elaborate on techniques to protect the privacy of users for certain machine learning tasks. Majeed and Lee [27] provide an overview of de-identification techniques for relational tables to complex social graphs. They classify the techniques into graph modification, generalization/clustering, privacy-aware graph computation, DP approaches and hybrid graph anonymity methods. Moreover, they come to the conclusion that traditional anonymization techniques do not perform well without further improvements. Rao et al. [33] compare de-identification techniques for large scale data in third party data sharing. They come to the conclusion that there is no concrete solution yet. Nevertheless, they see future potential in machine learning-based techniques. Wernke et al. [43] compare different privacy approaches to protect location privacy. They conclude that the combination of different attacks is still a challenge for the de-identification approaches they analyzed.

Rinaldo and Horeis [36] present a model to achieve a realistic assessment of autonomous structures considering the relation between safety and security. However, their approach does not consider privacy requirements at all.

3 METHODOLOGY

In the following section we explain how we have chosen, selected and evaluated de-identification techniques for our use case.

3.1 Scenario Development

The scenario was developed in multiple video calls with experts from the Research Association for Automotive Technology (FAT), a department of the German Association of the Automotive Industry (VDA[3]). Altogether, three scenarios were developed. The iterative procedure consisted of a presentation of the current version of the scenario by the authors of this paper. The presentation was intermingled and followed by feedback from the experts. After the feedback, the scenario was revised for the next presentation. Altogether, there were five feedback loops until the scenarios were considered to be mature.

3.2 Requirement Elicitation

To elicit the requirements in section 5, we started building a thread model for the presented use case (cf. section 4) in collaboration with the experts from FAT. From the related literature we identified potential risks and mapped them to the presented use case. We especially focus on risks for location-based sevices as, e.g. introduced by Wernke et al. [43]. A deeper discussion will exceed the focus of this paper because possible extension and mathematical definition have to be introduced. Again, with this paper we aim to provide a starting point to choose a de-identification technique when developing a location-based service in vehicular networks. Nevertheless, potential attacks and drawbacks of each technique are elaborated in the results (section 7).

After we identified the potential risks for the vehicle/driver we elicited the most important requirements the de-identification techniques have to fulfill.

From the scenario and the requirements we derive attributes to evaluate the de-identification techniques which meet the requirements.

3.3 Selection of De-identification Techniques

To identify possible techniques we use the ISO/IEC 20889:2018 and de-identification literature in the vehicular domain. To select suitable de-identification techniques we use the scenario definition and requirements. Insufficient solutions are also excluded.

3.4 Analysis of De-identification Techniques

Finally, we present possible implementations with the leftover de-identification techniques. We evaluate the de-identification techniques with the attributes derived earlier. Finally, we map the results in an overview table.

4 SCENARIO DESCRIPTION

The aim of this use case is to provide a third party, the Energy Grid Operator (EGO), with accurate and current weather data. This data is gathered by vehicles on the road within a specific area for which the EGO needs more or more accurate information. Figure 2 depicts a high-level overview of the use case while the entities in the use case are described below in more detail.

4.1 Entities

The entities in this use case are defined as follows:

- **E1 Vehicle** The vehicle driving within a certain geographical area is using multiple sensors to collect live weather

[3] German: *Verband der Automobilindustrie e. V.* is a German interest group of the German automobile industry consisting of automotive manufacturers as well as automobile component suppliers

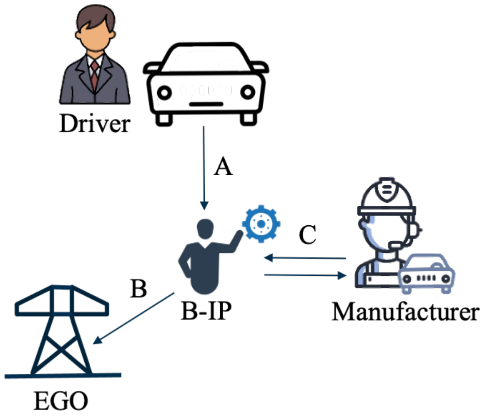

Figure 2: High-level data flow chart

information such as brightness, rain and humidity. This information, together with the current vehicle location, is used to provide real-time weather insights as a service to the B-IP. The concern of the driver is that no personal data about the driver of a vehicle is shared with the B-IP. The sensors of the vehicle create data in a frequency of 60 data points per minute. There are multiple vehicles on the road.

- **E2 Vehicle drivers** Vehicle drivers can be both the owner of a vehicle as well as other individuals such as friends and family members of the vehicle owner. For the technical evaluation of this use case, a differentiation is not necessary. For the sake of simplicity, we will also not differentiate between driver and vehicle.
- **E3 Business Intelligence Provider (B-IP)** The B-IP is responsible for analyzing and preparing the data and follows the need-to-know principle. Thus, the B-IP only receives data that is mandatory to meet the EGO's requirements.
- **E4 Electricity Grid Operator (EGO)** The EGO uses the data from the B-IP for energy demand predictions. Therefore, the EGO requires aggregated data once per minute. The data quality is required to exhibit enough information to make reliable energy demand predictions. Therefore, the EGO requires data in a frequency of 1 data point per minute.
- **E5 Manufacturer** The vehicle manufacturer is the initiator of the use case and receives information on the correct functioning of the service itself. This may include information on the total amount of data that has been processed and aggregated statistics on the provided service. The vehicle manufacturer is not directly involved in the processing and providing of data and is therefore not further considered in this scenario. No sensitive data is exchanged between B-IP and the manufacturer.

4.2 Data Flow

In this section we establish privacy sensitivity and data gathering frequency for the different types of data that are to be used in the use case. The different communication channels are derived from Figure 2. Table 1 is showing the data gathered by the vehicle and send to the B-IP via channel A.

Table 1: Communication Channel A: From vehicle to B-IP

Data	Privacy Sensitivity	Data truthfulness at record level	Frequency
Brightness	Low	No	1/min
Rain	Low	No	1/min
Temperature	Low	No	1/min
Atmospheric pressure	Low	No	1/min
Humidity	Low	No	1/min
GPS	High	No	1/min
VIN	High	Yes	1/min

Table 2: Communication Channel B: From B-IP to EGO

Data	Privacy Sensitivity	Data truthfulness at record level	Frequency
Brightness	Low	No	1/min
Rain	Low	No	1/min
Temperature	Low	No	1/min
Atmospheric pressure	Low	No	1/min
Humidity	Low	No	1/min
GPS	High	No	1/min

Both EGO and B-IP place requirements on the data quality and the frequency with which the data is to be provided to them. Table 2 is showing the data that arives at the B-IP and is send to the EGO.

4.3 Assumptions

This use case comes with several assumptions:

- **A1 Personal data sharing** No direct personal data is shared about the owners and drivers of a vehicle.
- **A2 De-Identification** An optimal solution provides data privacy for all data types that occur in this use case.
- **A3 Data frequency** For its purposes, the EGO requires the aggregated data once per minute.
- **A4 Data quality** The data quality after the data collection is sufficient for the EGO's purpose.

5 REQUIREMENT ELICITATION

Wernke et al. [43] claim that one approach for de-identification in location privacy scenarios is hiding the users' identity while only reveling the position of anonymous objects. One of the major threads they identified for this approach is the linking of context information with the anonymized location. Furthermore, according to them, another approach is to only provide location data to customers with a certain accuracy. Moreover, temporal information strongly influences the thread of context information linking [43].

5.1 Threat Model

In this section we present possible attack scenarios for the EGO use case. These form the basis for finding adequate solutions that can withstand such or similar attack scenarios. The thread model is derived from the protections goals of Wernke et al. [43] that covers: user identity, user position, and temporal information in combination with identity or position.

5.1.1 *Exact location determination* .

This attack tries to reveal the exact location of the vehicle from a perturbated GPS location. This is done by combining the perturbated GPS location with brightness or rain data and an external database containing tunnel data. If all cars in a certain area report and one car does not, one can assume that this car was driving through a tunnel at the time of reporting. Because the number of tunnels in a certain area is limited, the location can be guessed precisely and thereby the perturbation is annulated. Wernke et al. [43] describe this attack as map matching in which irrelevant areas are removed until a certain user can be identified.

5.1.2 Vehicle Tracking and Track Localization. Even with perturbated GPS signals, a malicious B-IP can easily be track certain vehicles if the speed limits on certain roads are known. This information can be easily accessed with an external database. Although the B-IP does not get the true location, the average speed can be calculated over time and based on this, possible roads or highways can be identified. Also, a database with traffic information containing traffic jams and accidents can leverage this attack. Gruteser and Grunwald [20] claim that privacy problems in vehicular environments are magnified if a service requires continous recording and sharing of location data.

5.1.3 Linkability and Profiling. If a VIN number can be clearly mapped to a certain vehicle, a malicious B-IP can easily profile a certain vehicle over time. Although data is sent perturbated and anonymized, the B-IP in this attack tries to identify certain vehicles and creates profiles over time.

5.2 Requirements for De-Identification

From the presented threads we derive the following requirements:

- Unlinkability: The B-IP should not be able to identify a certain vehicle to lower the risk of profiling. This also holds for the EGO who should also not be able to identify a certain vehicle from the crowd.
- Location perturbation: No real GPS data is sent to decrease the risk of identification. This requirement becomes more difficult over time and is closely related to linkability.
- The quality of data should still be high enough to add value to the EGO's energy demand prediction model.

5.3 Attributes for De-identification Techniques

Using the threat models derived in section 5, we identify suitable de-identification techniques that are able to provide an acceptable level of data privacy. For each technique, the level of effort and quality of results are determined. Hereby we focus on a qualitative evaluation based on academic literature without the use of real data. All de-identification techniques and the proposed de-identification techniques are evaluated on the following aspects that we derived from the requirements and the scenario:

- **Protective effect** The overall level of privacy that can be achieved through the de-identification technique in the particular use case. An optimal solution is able to protect personal information against any attack scenario outlined in this work.
- **Complexity** Complexity describes the overall complexity to develop, implement and maintain a particular solution. Oftentimes, a de-identification technique cannot simply be put to work but requires careful fine-tuning towards the specific type and frequency of gathered data as well as the desired output. Additionally, techniques and their algorithms need to interact with the environment in which they are implemented.
- **Runtime** Runtime describes the time that the overall solution for a use case needs to perform all necessary tasks that lead to the de-identification of data. This includes the actual runtime of algorithms, the execution of code, and the gathering and distributing of data and results between different entities.
- **Degree of maturity** The degree of maturity describes the scientific and commercial advancement of a de-identification technique. While some techniques are already used regularly, others need not yet be suitable for commercial use.
- **Implementation effort** The overall effort that needs to be taken to implement the solution for a specific use case. This includes the provision, installation and fine-tuning of hardware and software for the specific entities as well as the time and human resources that are needed for its implementation.
- **Monetary cost** Monetary cost includes the cost of development and procurement of all necessary hard- and software for each use case. Possible interfering factors are use case-specific circumstances and factors that might hinder the performance, effectiveness and efficiency of the de-identification technique.

The quality of the data after the implementation of the de-identification techniques are evaluated on the following aspects:

- **Time blur** Time blur depicts the degree to which data loses information that are related to a specific time point. That means data that is gathered over a period of time and might then be aggregated to a single data point. Here, time-related information gets lost, resulting in time blur.
- **Time delay** Time delay depicts the delay with which data is reported and can be acted upon. That is, data might be collected continuously but loses its value as the computation of results takes significant time, resulting in a time delay that decreases the value of created insights.
- **Location obfuscation** Location obfuscation depicts the degree of obfuscation applied in a specific scenario. Location data might for instance be aggregated on a street, city or kilometer basis.
- **Processing speed** describes the execution time of the de-identification technique itself.
- **Aggregated data** Aggregated data describes a state in which data that is gathered during a use case is aggregated and thus a loss of information in the data occurs. While most scenarios allow for some aggregation, as the amount of data that is produced is high, more aggregation is likely to decrease the usability of a de-identification technique.
- **Truthfulness** Truthfulness describes whether input data and output data are equal when using a de-identification technique. Different techniques may report non-truthful data when data is perturbed, noise is added or the sequence of data is changed. Less truthful data output can decrease validity of insights that are generated in a use case. The combined

evaluation of the different aspects described above enables us to make a statement on the overall suitability and usability of a de-identification technique for a particular use case. For each use case, a table is provided that compares all suitable de-identification techniques against each other. Factors are ranked as Low, Medium and High, whereby a color-code using red, yellow and green demonstrates the positive or negative effect of that ranking. For instance, a technique may score High on complexity, which would result in a red color-code, as a high degree of complexity is not seen as favorable.

6 SELECTION OF DE-IDENTIFICATION TECHNIQUES

We aim to evaluate the de-identification techniques identified in section 2.1 upon the EGO use case. Thus, this section includes the technical evaluation of suitable de-identification techniques.

Upon accessing the assumption and requirements of the electricity provider use case, all de-identification methods introduced in 2.1 have been evaluated for their fit for the use case. Only de-identification methods that could initially demonstrate a sufficient level of privacy are discussed in detail below. In general, Wernke et al. [43] focus on three dimensions to evaluate de-identification techniques: user identity, user position and identity/position in combination with time. They claim that a common approach for the de-identification in location privacy scenarios is hiding the users' identity while only reveling the position of anonymous objects. One of the major threads they identified for this approach is the linking of context information with the anonymized location. Furthermore, they claim that to keep the users position secret, location data to customers should only provided with a certain accuracy. From their point of view, temporal information strongly influences the thread of context information linking [43]. In the following, we depict methods that have been excluded, as well as a brief statement on as to why they are deemed not suitable in this particular use case.

Sampling does not provide privacy protection for the subset of data and relies on a very high sample size, which is likely not to be the case for vehicles driving in rural areas.

Considering the *cryptographic approaches*, deterministic encryption is not suitable for weather data, as only a very limited set of information will be used. This makes re-identification possible. Order-preserving encryption is also not suitable for our use case because the order of data is not important. In contrast to that, HE can provide the B-IP with information about the weather in a certain while, the data itself of a certain vehicle remains encrypted. Therefore, we include HE for further consideration. Also SMC that can be treated as a cryptographic approach fulfills the requirements of section 5.

In general, *suppression* does not fulfill the requirements because removing certain values will significantly decrease the data quality and therefore violate requirements [23]. Nevertheless, suppression can be very well used in combination with other techniques. For example, removing direct identifiers of vehicles and removing unique values are essential for some of the later mentioned de-identification approaches.

Similar to supression, *pseudonymization* would only work for identifiers such as a vehicle ID in our dataset. A pseudonymization of e.g. location or temperature will have a strong negative impact on the data quality.

Generalization is suitable for sensor data with the data type float, e.g., brightness, rain or temperature data. While rounding is very easy to implement, top/bottom coding lags in a useful definition of a threshold for the weather data. Moreover, rounding and top/bottom coding are too weak for state-of-the-art GPS de-identification or make the data unusable for location-based services [23].

Similar to the techniques above, *randomization* alone does not protect against location tracking or corrupts the data in a way that it becomes useless for location-based services [23] . Also, the permutation of data does not work for a trajectory. The route of a vehicle could still be identified.

TEE require a trusted third party for the setup of the TEE. Since that body would be considered to be the data controller, this results in several problems. First, it is unclear how assurances that the TEE fulfills its purpose could be conveyed to the driver. Second, the legal classification of responsibility for the TEE is not yet fully clarified, and thus it remains unclear if data processing within the TEE can contribute to the de-identification of data. TEEs seem to be more appropriate to guarantee the correctness and freshness of the data. However, a full assessment of TEEs is beyond the scope of this paper.

DP must be considered in several dimensions. While central DP allows the B-IP to see the data before the de-identification, local DP might exhibit problems in the frequency of data sharing. Nevertheless, the Encryption Shuffle Analyze (ESA) architecture proposed by Bittau et al. [8] is able to overcome these issues. Also location perturbation with local DP as presented by [5] is useful for the use case.

K-anonymity is a comparatively simple concept that is easy to implement, utilizing techniques such as data suppression or generalization to create a k-anonymous dataset. Moreover, k-anonymity provides a trade-off between usability and privacy so the level of de-identification and data quality have to be evaluated independently for each scenario [6, 42].

In *FL*, the data of each vehicle stays local and only the local models' gradients are shared with the B-IP. In theory it is possible to design a model that fulfills the requirements.

To put it in a nutshell, the de-identification techniques that are evaluated to be initially suitable are HE, MPC, Distributed DP, FL and k-anonymity.

7 ANALYSIS OF DE-IDENTIFICATION TECHNIQUES

In this section we analyze and compare the different privacy preserving data analysis approaches identified as suitable in the previous section.

7.1 Homomorphic Encryption

As explained in previous chapters the advantage of HE is that data can be computed while it is encrypted, guaranteeing that computations on the data lead to the same results on the decrypted data.

In the HE de-identification approach, the electricity provider distributes a secret key to the vehicles on the road (see Figure 3). The vehicles that collect the weather data then use their key to homomorphically encrypt their location and weather data. The data is then distributed to the B-IP. The B-IP is now able to process the data as determined by the electricity provider. Meta data is deleted and average weather and location data is sent to the electricity provider. All these operations are performed on encrypted data, the B-IP is therefore unable to gain insights into vehicles actual locations and other provided information. However, operations on the decrypted data result in the same operation on the underlying data. The electricity provider is now able to use its secret key to decrypt the data and use the encrypted results for the intended purpose.

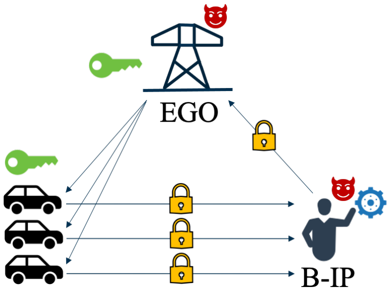

Figure 3: HE data flow chart

Nonetheless, HE creates several drawbacks. Although the technique itself has been available for some time, its actual usefulness is still hindered through the loss of performance and computational speed. Only a limited number of different operations, e.g. addition and subtraction can be computed, while runtime increases greatly with the number of computations. However, research on homomorphic algorithms continues to improve runtime, making HE a suitable solution for mobility-related use cases in the near future. Additionally, the techniques do not rely on the number of vehicles on the road and do not decrease the actual usability and truthfulness of the data.

7.2 Secure Multiparty Computation

MPC can be realized using a map that is separated in different clusters e.g., with a grid (see Figure 4). Vehicles in each of these clusters calculate the average energy demand of a certain cluster with secure multiparty computation. One vehicle of each cluster is then chosen as the cluster leader that sends the computed result to the B-IP. To avoid an identification of certain vehicles by the B-IP shuffling between the cluster leaders is also possible. In general, this can also have a positive effect on the minimum cluster size because it can be larger if the vehicles cannot be linked to a certain cluster. Nevertheless, the more vehicles in a cluster, the better the accuracy of the weather infomation and the better the de-identification because a single vehicle can can be hidden more easily in the crowd. Finally, the B-IP receives only the average cluster data. Based on this data, e.g. a heatmap with the weather information can be derived that is then shared with the EGO.

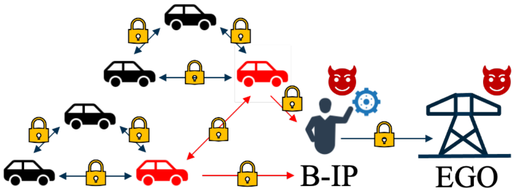

Figure 4: MPC data flow chart

One approach for vehicular MPC communication is provided by Li et al. [42] who propose a cooperative control strategy incorporating with efficient MPC, reducing latency and integrating a function secret sharing scheme.

First, one interfering factor for this de-identification technique is the vehicle density required per cluster. In case not enough vehicles are located in a certain cluster, no information can be calculated and sent to the B-IP. Second, a stable connection between the cars is required to use the MPC protocol. Third, the communication between the vehicles is likely to produce a huge overhead so that besides a good network coverage, a minimum bandwidth is mandatory.

7.3 *K*-anonymity

K-anonymity in itself is not a de-identification technique but a property with which data privacy in a database might be measured. ISO/IEC 20889:2018 defines *k*-anonymity as a formal privacy measurement model which ensures that an equivalence class in a database contains at least K records that are similar for each identifier.

In the EGO use case, the objective of vehicles is to obfuscate their exact location and ensure that weather information cannot be used for location inference. For *k*-anonymity, a map is clustered into various mix points whereby each mix point fulfills the *k*-anonymity requirement (see Figure 5). In the electricity provider use case, the map represents the area in which vehicles are to gather weather information. This area is divided into mix points to increase the accuracy of information. The work of Corser et al. [13] introduces multiple different protocols to create mix points such as stationary mix points, mix points occurring at irregular time intervals or randomly chosen mix points that may occur regularly or irregularly. An additional option would be that vehicles themselves create mix points and act as group leaders of other vehicles, thereby managing the fulfillment of *k*-anonymity and the data distribution behavior of a group of vehicles. Within such a mix point, whose center may for instance be an intersection, vehicles switch pseudo IDs with other vehicles and/or are added to an anonymity set and do not communicate information for a specific time period. Essentially,

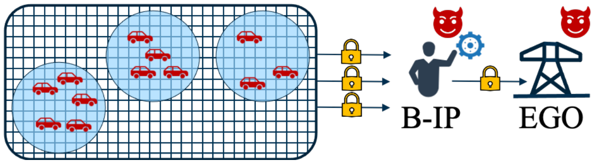

Figure 5: *K*-anonymity data flow chart

the work uses the de-identification techniques of suppression and pseudonymization to achieve k-anonymity. Additionally, such a model could be enhanced by adding further simple de-identification techniques such as aggregation, noise addition or permutation to it. Such options would enhance privacy at the cost of a loss of quality of service as the usefulness of data decreases. In [14] the authors decided against the use of such options as anonymizing, for instance through spatial cloaking, cannot effectively protect against tracking over time and leads to less precise results. Dummifying has not been used as false location data might lead to accidents as the authors' use case has been to provide relevant safety traffic data to other vehicles through a central service. However, in our use case, exact location data is not as important as in other use cases as the weather might not differ strongly in a 500m radius. Time delay might also be acceptable to an extent, as weather will not change significantly within 5 minutes.

Therefore, a combination of simple de-identification techniques that fulfill k-anonymity are seen as a suitable alternative for the EGO use case. In any case, the protective effect of this solution will not be as high as that of more advanced methods such as MPC. Multiple factors affect the level of privacy that can be obtained: A lower vehicle density results in a lower K-value and a lower level of privacy. The topology, e.g. the number of roads and the speed of travel, influence privacy as fewer roads lead to less privacy. Similarly, the choice and design of mixing points, depending on the chosen protocol, need to be matched with such factors.

Complexity of the model is low while the runtime again depends on the choice of techniques and protocols used. Such protocols however already exist, creating mature solutions that could be implemented quickly and at low monetary cost. As elaborated, data may be sent from each vehicle or aggregated between vehicles. Data could include dummy variables, resulting in non-truthful data. Depending on the number of vehicles in a mix point and on the road, the usefulness of the data might change. Less vehicles equal larger mixing points and an increase in location obfuscation and possibly time delay in order to ensure privacy.

Overall, while k-anonymity-based solutions might provide a cheap solution that can be implemented easily, data quality and the achievable level of privacy greatly depend on topology and the number of vehicles within an area.

7.4 Distributed Differential Privacy

For this de-identification technique we utilize the system architecture Encode, Shuffle, Analyse (ESA) proposed by Bittau et al. [8] to implement distributed DP (see Figure 6). In general, the architecture consists of three entities, an encoder, a shuffler and an analyzer, as seen above. In the following we will have a detailed look at the tasks of each entity in our concrete scenario with the EGO.

Encoder: The encoder is responsible for ensuring the fulfillment of the user's trust assumptions by locally transforming and conditioning the user's private data [8]. In our EGO use case one of these transformations is the location perturbation providing local DP as proposed by Andrés et al. [5] by fulfilling the requirement of geo-indistinguishability. Moreover, the encoder is responsible for the encryption of the data with an inner and outer encryption, and the transmission over a secure channel to the shuffler. As explained above, the encoder entity is placed on the user's device, in the EGO use case, we place the encoder in the car.

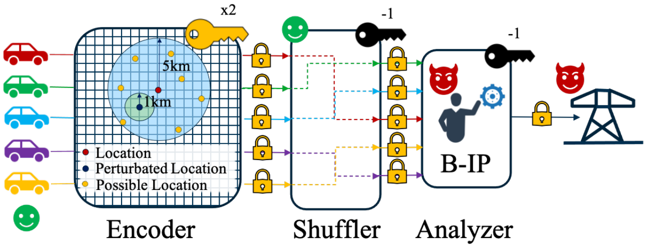

Figure 6: DP data flow chart

Shuffler: The shuffler acts as an additional privacy layer in between the user's encoder and the analyzer that should be run by a trusted third party. The shuffler is responsible for the anonymization, shuffling, thresholding, and batching of the data received from the encoder. By decrypting the outer encryption, the shuffler can access the metadata of a user, e.g., timestamps, source IP addresses, routing paths. The main task of the shuffler is to remove all this data before forwarding it to the analyzer. To prevent the reassignment of the data by the analyzer to a certain user, the data are reordered randomly and forwarded infrequently and only in batches. Moreover, the shuffler can also set thresholds and reject data items to ensure that each item can be hidden in a sufficient crowd.

Analyzer: The analyzer is responsible for the innermost decryption, storing and aggregation of the data received from the shuffler. The analyzer utilizes techniques such as DP to make the data available for other groups of interest without revealing private user information. In the EGO use case this role is taken by the B-IP. The B-IP uses the data received from the shuffler to create a weather map that is sent to the EGO.

The biggest issue of this approach is car density and appears if only viewed cars are in a certain location. As a result of this, a single car cannot be hidden sufficiently in the crowd and the shuffler has to delay or withdraw the forwarding of certain batches. Therefore, a minimum number of cars per region is required. Moreover, the number of cars is influenced by area topology and daytime. In a scenario where the EGO wants to make assumptions on the required network load, e.g. for vehicular charging, the absence of data in a certain region would point to a very low electricity demand. The average demand for an area could be set approximately on historic results or in dependency of the minimum number of cars.

7.5 Federated Learning

Similar to the MPC framework, the FL as de-identification technique can be realized by dividing the map in grid-based clusters. The vehicles in the clusters then share data with each other (see Figure 7). To ensure de-identification in vehicle-to-vehicle communication, a MPC protocol can be utilized. Besides the grid approach, the clusters can also be determined utilizing the cars' communication radius similar to the approach of Yin et al. [45]. In both cases, the cars will have to communicate with each other to determine a leader of each cluster. In case the weather parameters in a certain cluster did not change, the leader will not participate in the current round of training to keep the traffic as low as possible. To fulfill

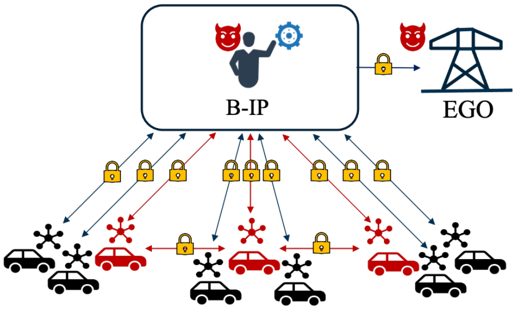

Figure 7: FL data flow chart

the requirement of unlikability, a distributed shuffling protocol as proposed by Cheu et al. [12] between all Leaders of a cluster can be used to delete metadata and shuffle the data between the Leaders. Location perturbation of the leaders within the clusters could also be helpful. The leaders participating in a training round are then responsible to send the locally derived model to the B-IP. The B-IP cannot link the received data to the sender because the data was shuffled before and metadata was deleted. In each training round, e.g., every minute, the B-IP receives updated models from the leaders. These models are then used to develop a new central model. This model is then sent to the EGO and also distributed to all cars. A possible extension to keep the traffic low is to determine the new Leader for a certain round in advance and only use the Leader's data in that round. The Leader could still exchange data with other cars in the cluster but the model is only with the Leader. In future, further experiments are required to build the most efficient model.

In literature some similar approaches were already proposed. Saputra et al. [39] propose a FL model for energy demand prediction for electric vehicle networks, but compared to our approach they utilize the information gathered from the charging stations. On the one hand, they use a FL model with the aim to reduce the communication overhead between the charging stations and the main server with the central server. On the other hand, they protect the data of the vehicle users by only transmitting relevant information in the form of parameter updates to the central server rather than sending whole data sets. Liu et al. [26] present a traffic flow prediction scheme using location-based clustering in combination with a FL approach. In their approach they collect the information from organizations (e.g., bus stop or station) while randomly selecting only a defined ratio of organizations from a larger group in each round of training. Yin et al. [45] propose a Federated Localization (FedLoc) framework with the aim to build accurate location services without revealing sensitive user information. They propose a cloud-based network infrastructure that is based on many clusters that do not overlap. These clusters are defined by the mobile communication range of a mobile terminal, e.g., 5G macro and micro base stations and WiFi6-networks that can enable a high communication rate.

The biggest pitfalls for the FL approach are vehicle density and network coverage. A minimum number of vehicles is required to form a cluster, otherwise no information can be sent to the B-IP. Moreover, a lot of communication is required for this distributed learning approach, therefore a sufficient network coverage is mandatory.

7.6 Comparison of De-Identification Technologies

In Table 3 we provide an overview of the attribute-evaluation for all de-identification techniques. These results are derived from the de-identification technique specific analysis. In Table 4 we summarize all interfering factors. It is important to mention that an "x" only indicates that the de-identification technique is sensitive to small occurrences of this interfering factor. For larger occurrences, all de-identification techniques are effected. For example, if there is only one vehicle, all de-identification techniques will struggle.

Table 3: Aggregated results of de-identification techniques

	HE	SME	Distr. DP	FL	K-anon.
Protective effect	⊕ High	⊕ High	⊕ High	⊕ High	⊙ Medium
Complexity	⊖ High	⊖ High	⊖ High	⊖ High	⊕ Low
Runtime	⊖ High	⊖ High	⊖ High	⊖ High	⊙ Medium
Degree of maturity	⊙ Medium	⊙ Medium	⊕ High	⊙ Medium	⊕ High
Implement. effort	⊖ High	⊖ High	⊖ High	⊖ High	⊕ Low
Monetary cost	⊙ Medium	⊖ High	⊖ High	⊖ High	⊕ Low
Time blur	⊖ High	⊙ Medium	⊕ Low	⊙ Medium	⊙ Medium
Location obfus.	⊕ Low	⊖ High	⊕ Low	⊕ Low	⊖ High
Processing speed	⊖ Low	⊖ Low	⊕ High	⊕ High	⊙ Medium
Time delay	⊙ Medium	⊖ High	⊙ Medium	⊙ Medium	⊙ Medium
Aggregated data	Yes	Yes	Yes	Yes	Yes
Truthfulness	Yes	No	Yes[1]	Yes	No

[1] (No for GPS)

Table 4: Possible interfering factors

	HE	SME	Distr. DP	FL	K-anon.
Communic. overhead	x	x		x	
Network coverage	x	x		x	x
Area topology			x		x
Car density		x	x	x	x
Car speed	x	x	x	x	x

8 DISCUSSION

In this section we provide a better understanding of the results, including impact, limitations and future work.

8.1 Impact

When comparing the different solutions for the electronic grid operator use case, we find that all advanced de-identification techniques are able to provide a high level of privacy for individuals and vehicles. However, all solutions are relatively complex and most of them require further research or an extension to mitigate some of the drawbacks, such as communication overhead or computational costs. Although the de-identification techniques are very different, they all exhibit the trade-off between usability of the data provided to the EGO and the de-identification of the vehicle/driver. In practice, this trade-off will be complicated by specific project restrictions such as costs, project duration or expected service lifetime.

For example, while a solution based on k-anonymity offers the least amount of privacy protection, it is easily implementable, cheap with an acceptable data output for the EGO. On the one hand, distributed DP and FL are both more complex solutions, but on the other hand, they provide more fine-grained insights as the data quality remains higher. Very accurate results can also be achieved with HE, but the calculation on encrypted data might be slow and require much more resources. Another drawback occurs if every vehicle communicates directly with the B-IP. As with FL this can be compensated by e.g. implementing data processing at the edge that aggregate results before they are sent to the B-IP for further processing. Nevertheless, this has also drawbacks because the complexity of the network typology will further increase and reveal more targets for attacks. In general, distributed de-identification techniques like FL have the advantage that data is processed directly on the device. This decreases communication overhead and the computational effort at the B-IP. The B-IP can also not be affected by hacking attacks in which large amounts of data are stolen because such data does simply not exist. Nevertheless, FL is a relatively new technology, so the absence of know-how might highly influence the decision which technology to choose.

The external factors such as vehicle density, traveling speed and network coverage that we identified during our analysis for each de-identification technique are likely to significantly influence the stable execution of each use case. Therefore, these problems should be considered as systematic risk to the scenario that requires some effort to be compensated. For example, traffic flow simulations could be used to verify solutions by combining simulated traffic scenarios with actual vehicle data.

In the scenario description we defined the vehicle manufacturer as a passive entity that monitors the scenario. In practice, the manufacturer will initiate more than one service, and some of these services will also require the transmission of sensitive data. This is the reason why we have not excluded the manufacturer from the beginning.

8.2 Limitations

One of the major pitfalls of the proposed de-identification techniques is the theoretical approach that was used to evaluate the techniques. Although advantages and disadvantages of each techniques were identified, they should be understood more as a general guideline. For example, the real performance of a technique can be only tested in practice using real vehicular data and including all inferring factors that have impact on data quality, delay of the service, effort and costs.

Although not considered in the analysis so far, the knowledge and past experience with a certain de-identification technique in the implementing body can have a huge impact on the cost decision and implementation effort.

Another pitfall to consider is the legal assessment of each evaluation technique. For example, HE is a key-based approach. Although the keys are kept secret, there is the chance to steal the vehicle's key. Also, the privacy guarantee off DP is only a mathematical construct and not a standardized method. Including different entities, such as the shuffler, and minimum batch sizes, the underlying mathematical construct changes or can, in the worst case, only be approximated.

8.3 Future Work

As mentioned in the limitations, the analysis of de-identification techniques is missing a technical approach to identify possible pitfalls during the implementation. In the future, we will work on a technical comparison, e.g. using simulation or in depth technical evaluation. Moreover, while machine learning is becoming more relevant and the computation on client devices solves the problem of communication overhead and data leakage, more work with the focus on de-identification approaches for distributed learning techniques should be carried out.

9 SUMMARY AND CONCLUSION

In this paper we provide clarity on the relevant techniques for the de-identification of location-based services in the automotive area. Focusing on the demand of developers with a similar scenario in particular, we aim to provide decision support for the selection of a suitable de-identification technique. To achieve this, we have analyzed the third party vehicular data sharing using the example of the EGO scenario. For this scenario we have identified the privacy threats that are: exact location determination, vehicle tracking and track localization, and linkability and profiling. Based on these threats, we elicited requirements that helped us to select de-identification techniques from ISO/IEC 20889:2018 and further literature on vehicular de-identification techniques. We identified the 5 techniques homomorphic encryption, secure multiparty computation, distributed differential privacy, federated learning and k-anonymity, which we compared on the basis of a number of relevant attributes. We find that no strategy is dominating the others because all techniques provide an increased privacy but differ strongly in the other attributes. Our contribution is the elaboration of these attributes per technique. Based on our scenario, we provided possible topologies of the de-identification techniques and explained the relation and communication between the related entities. We also analyzed the potential computational effort of each entity and possible pitfalls for the de-identification techniques.

Our evaluation of de-identification techniques has shown that within each de-identification technique different approaches to calculate the privacy gain. E.g., for differential privacy, standardized methods would make the comparison of techniques much easier. We also find that most de-identification techniques highly depend on the network coverage. With a high bandwidth and stable connections, the bottleneck of communication overhead can also be reduced. Moreover, we conclude that to keep costs low, privacy has to be considered from the beginning to ensure that the offered service is at the same time efficient and privacy preserving.

Finally, our results built a starting point to choose a sufficient de-identification technique for a vehicular location data sharing scenario. As our evaluation of the attributes for the de-identification techniques is only based on the current literature, we will plan a technical evaluation in the next steps.

ACKNOWLEDGMENTS

The authors are grateful to the Forschungsvereinigung Automobiltechnik e.V. (FAT e.V.) who funded this research. We are in particular grateful to FAT's working group "AK 31 Elektronik und Software" who not only initiated this research and provided input and feedback on the underlying report [32] in various meetings.

REFERENCES

[1] 2014. Consumer Privacy Protection Principles – PRIVACY PRINCIPLES FOR VEHICLE TECHNOLOGIES AND SERVICES. https://cryptome.org/2014/11/auto-privacy-principles.pdf

[2] 2017. Vehicle Data Privacy – Industry and Federal Efforts Under Way, but NHTSA Needs to Define Its Role. https://www.gao.gov/assets/gao-17-656.pdf

[3] 2020. Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications. https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202001_connectedvehicles.pdf

[4] Mohammad Al-Rubaie and J Morris Chang. 2019. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy* 17, 2 (2019), 49–58.

[5] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the ACM Conference on Computer and Communications Security*. https://doi.org/10.1145/2508859.2516735 arXiv:1212.1984

[6] J. Andrew, J. Karthikeyan, and Jeffy Jebastin. 2019. Privacy Preserving Big Data Publication On Cloud Using Mondrian Anonymization Techniques and Deep Neural Networks. In *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*. 722–727. https://doi.org/10.1109/ICACCS.2019.8728384

[7] Michele Bertoncello, Gianluca Camplone, Paul Gao, Hans-Werner Kaas, Detlev Mohr, Timo Möller, and Dominik Wee. 2016. Monetizing car data—new service business opportunities to create new customer benefits. *McKinsey & Company* (2016).

[8] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. 2017. PROCHLO: Strong Privacy for Analytics in the Crowd. In *SOSP 2017 - Proceedings of the 26th ACM Symposium on Operating Systems Principles*. https://doi.org/10.1145/3132747.3132769 arXiv:1710.00901

[9] Christoph Buck and Riccardo Reith. 2020. Privacy on the road? Evaluating German consumers' intention to use connected cars. *International Journal of Automotive Technology and Management* 20, 3 (2020), 297–318.

[10] Alexandra Campmas, Nadina Iacob, Felice Simonelli, and Hien Vu. 2021. Big Data and B2B platforms: the next big opportunity for Europe – Report on market deficiencies and regulatory barriers affecting cooperative, connected and automated mobility.

[11] Valerie Chen, Valerio Pastro, and Mariana Raykova. 2019. Secure computation for machine learning with SPDZ. *arXiv preprint arXiv:1901.00329* (2019).

[12] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. 2019. Distributed differential privacy via shuffling. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-17653-2_13 arXiv:1808.01394

[13] George P Corser, Huirong Fu, and Abdelnasser Banihani. 2016. Evaluating location privacy in vehicular communications and applications. *IEEE transactions on intelligent transportation systems* 17, 9 (2016), 2658–2667.

[14] George P. Corser, Huirong Fu, and Abdelnasser Banihani. 2016. Evaluating Location Privacy in Vehicular Communications and Applications. *IEEE Transactions on Intelligent Transportation Systems* 17, 9 (2016), 2658–2667. https://doi.org/10.1109/TITS.2015.2506579

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/11681878_14

[16] European Parliament and Council of The European Union. 2016. REGULATION (EU) 2016/679 General Data Protection Regulation (GDPR). http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=DE

[17] Fifi Farouk, Yasmin Alkady, and Rawya Rizk. 2020. Efficient privacy-preserving scheme for location based services in vanet system. *IEEE Access* 8 (2020), 60101–60116.

[18] Sebastian Frank and Arjan Kuijper. 2020. Privacy by Design: Survey on Capacitive Proximity Sensing as System of Choice for Driver Vehicle Interfaces. In *Computer Science in Cars Symposium*. 1–9.

[19] Michael Gardiner, Alexander Truskovsky, George Neville-Neil, and Atefeh Mashatan. 2021. Quantum-safe Trust for Vehicles: The race is already on. *Queue* 19, 2 (2021), 93–115.

[20] Marco Gruteser and Dirk Grunwald. 2003. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*. 31–42.

[21] ISO/IEC 20889:2018. 2018. Privacy enhancing data de- identification terminology and classification of techniques. *INTERNATIONAL STANDARD* (2018).

[22] Ioannis Krontiris, Kalliroi Grammenou, Kalliopi Terzidou, Marina Zacharopoulou, Marina Tsikintikou, Foteini Baladima, Chrysi Sakellari, and Konstantinos Kaouras. 2020. Autonomous Vehicles: Data Protection and Ethical Considerations. In *Computer Science in Cars Symposium*. 1–10.

[23] John Krumm. 2007. Inference attacks on location tracks. In *International Conference on Pervasive Computing*. Springer, 127–143.

[24] Atul Kumar, Manasi Gyanchandani, and Priyank Jain. 2018. A comparative review of privacy preservation techniques in data publishing. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. IEEE, 1027–1032.

[25] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* (2020). https://doi.org/10.1109/MSP.2020.2975749 arXiv:1908.07873

[26] Yi Liu, James J.Q. Yu, Jiawen Kang, Dusit Niyato, and Shuyu Zhang. 2020. Privacy-Preserving Traffic Flow Prediction: A Federated Learning Approach. *IEEE Internet of Things Journal* (2020). https://doi.org/10.1109/JIOT.2020.2991401 arXiv:2003.08725

[27] Abdul Majeed and Sungchang Lee. 2020. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access* (2020).

[28] Suntherasvaran Murthy, Asmidar Abu Bakar, Fiza Abdul Rahim, and Ramona Ramli. 2019. A comparative study of data anonymization techniques. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 306–309.

[29] Sebastian Pape and Kai Rannenberg. 2019. Applying Privacy Patterns to the Internet of Things' (IoT) Architecture. *Mobile Networks and Applications (MONET) – The Journal of SPECIAL ISSUES on Mobility of Systems, Users, Data and Computing* 24, 3 (06 2019), 925–933. https://doi.org/10.1007/s11036-018-1148-2

[30] Mert D Pesé and Kang G Shin. 2019. Survey of Automotive Privacy Regulations and Privacy-Related Attacks. (2019).

[31] Gunasekaran Raja, Sudha Anbalagan, Geetha Vijayaraghavan, Priyanka Dhanasekaran, Yasser D. Al-Otaibi, and Ali Kashif Bashir. 2020. Energy-Efficient End-to-End Security for Software Defined Vehicular Networks. *IEEE Transactions on Industrial Informatics* 3203, c (2020), 1–1. https://doi.org/10.1109/tii.2020.3012166

[32] Kai Rannenberg, Sebastian Pape, Frederic Tronnier, and Sascha Löbner. 2021. *Study on the Technical Evaluation of De-Identification Procedures for Personal Data in the Automotive Sector*. Technical Report. Goethe University Frankfurt. https://doi.org/10.21248/gups.63413

[33] P Ram Mohan Rao, S Murali Krishna, and AP Siva Kumar. 2018. Privacy preservation techniques in big data analytics: a survey. *Journal of Big Data* 5, 1 (2018), 1–12.

[34] Devin Reich, Ariel Todoki, Rafael Dowsley, Martine De Cock, and Anderson CA Nascimento. 2019. Privacy-preserving classification of personal text messages with secure multi-party computation: An application to hate-speech detection. *arXiv preprint arXiv:1906.02325* (2019).

[35] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. 2016. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication* 47 (2016), 131–151.

[36] Rhea C Rinaldo and Timo F Horeis. 2020. A Hybrid Model for Safety and Security Assessment of Autonomous Vehicles. In *Computer Science in Cars Symposium*. 1–10.

[37] Christian Roth, Sebastian Aringer, Johannes Petersen, and Mirja Nitschke. 2020. Are sensor-based business models a threat to privacy? the case of pay-how-you-drive insurance models. In *International Conference on Trust and Privacy in Digital Business*. Springer, 75–85.

[38] P Samarati and L Sweeney. 1998. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion. *Proc of the IEEE Symposium on Research in Security and Privacy* (1998).

[39] Yuris Mulya Saputra, DInh Thai Hoang, DIep N. Nguyen, Eryk Dutkiewicz, Markus Dominik Mueck, and Srikathyayani Srikanteswara. 2019. Energy demand prediction with federated learning for electric vehicle networks. In *2019 IEEE Global Communications Conference, GLOBECOM 2019 - Proceedings*. https://doi.org/10.1109/GLOBECOM38437.2019.9013587

[40] Andreas Tomandl, Florian Scheuer, and Hannes Federrath. 2012. Simulation-based evaluation of techniques for privacy protection in VANETs. In *2012 IEEE 8th international conference on wireless and mobile computing, networking and communications (WiMob)*. IEEE, 165–172.

[41] Jinbao Wang, Zhipeng Cai, and Jiguo Yu. 2019. Achieving personalized k-Anonymity-Based content privacy for autonomous vehicles in CPS. *IEEE Transactions on Industrial Informatics* 16, 6 (2019), 4242–4251.

[42] Jinbao Wang, Zhipeng Cai, and Jiguo Yu. 2020. Achieving Personalized k-Anonymity-Based Content Privacy for Autonomous Vehicles in CPS. *IEEE Transactions on Industrial Informatics* 16, 6 (2020), 4242–4251. https://doi.org/10.1109/TII.2019.2950057

[43] Marius Wernke, Pavel Skvortsov, Frank Dürr, and Kurt Rothermel. 2014. A classification of location privacy attacks and approaches. *Personal and ubiquitous computing* 18, 1 (2014), 163–175.

[44] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology* (2019). https://doi.org/10.1145/3298981

[45] Feng Yin, Zhidi Lin, Yue Xu, Qinglei Kong, Deshi Li, Sergios Theodoridis, and Shuguang Cui. 2020. FEDLOC: Federated learning framework for data-driven cooperative localization and location data processing. https://doi.org/10.1109/ojsp.2020.3036276 arXiv:2003.03697

[46] Liane Yvkoff. 2020. The Success Of Autonomous Vehicles Hinges On Smart Cities. Inrix Is Making It Easier To Build Them. Forbes. https://www.forbes.com/sites/lianeyvkoff/2020/10/28/the-success-of-autonomous-vehicles-hinges-on-smart-cities-inrix-is-making-it-easier-to-build-them/