Introducing Differential Privacy to the Automotive Domain: Opportunities and Challenges

Introducing Differential Privacy to the Automotive Domain: Opportunities and Challenges

Boel Nelson and Tomas Olovsson
Department of Computer Science and Engineering
Chalmers University of Technology, Sweden
Email: {boeln, tomasol}@chalmers.se

***Abstract*—Privacy research is attracting increasingly more attention, especially with the upcoming general data protection regulation (GDPR) which will impose stricter rules on storing and managing personally identifiable information (PII) in Europe. For vehicle manufacturers, gathering data from connected vehicles presents new analytic opportunities, but if the data also contains PII, the data comes at a higher price when it must either be properly *de-identified* or gathered with contracted consent from the drivers.**

One option is to establish contracts with every driver, but the more tempting alternative is to simply de-identify data before it is gathered, to avoid handling PII altogether. However, several real-world examples have previously shown cases where *re-identification* of supposedly anonymized data was possible, and it has also been pointed out that PII has no technical meaning. Additionally, in some cases the manufacturer might want to release statistics either publicly or to an original equipment manufacturer (OEM). Given the challenges with properly de-identifying data, structured methods for performing de-identification should be used, rather than arbitrary removal of attributes believed to be sensitive.

A promising research area to help mitigate the re-identification problem is *differential privacy*, a privacy model that unlike most privacy models gives mathematically rigorous privacy guarantees. Although the research interest is large, the amount of real-world implementations is still small, since understanding differential privacy and being able to implement it correctly is not trivial. Therefore, in this position paper, we set out to answer the questions of how and when to use differential privacy in the automotive industry, in order to bridge the gap between theory and practice. Furthermore, we elaborate on the challenges of using differential privacy in the automotive industry, and conclude with our recommendations for moving forward.

I. Introduction

The ability to collect data from modern connected vehicles presents opportunities for increased analysis, which enables vehicle manufacturers to both improve existing as well as develop new services. For example, investigating driving behaviour would make it possible to learn more about the drivers' needs and preferences, allowing manufacturers to better cater to customers' needs. Especially, using machine learning on large data sets could result in interesting correlations that were previously unknown.

However, gathering data from vehicles is not only an opportunity for further analysis, but also a possible privacy risk to the individual drivers. A recent survey show that drivers' privacy concerns include disclosure of private information, car vehicle tracking and commercial use of their personal data [1]. Seeing as privacy is a concern for drivers when it comes to connected vehicles, the problem needs to be addressed by the manufacturers in order to maintain the drivers' trust. Moreover, the upcoming general data protection regulation (GDPR) [2] will soon enforce stricter handling of personally identifiable information (PII). Failure to comply with the GDPR may result in fines of up to either €20,000,000 or 4% of the total worldwide annual turnover of the preceding financial year [2]. Even though the GDPR is a European law, it will affect all companies that sell vehicles to Europe, as this is where the data will be collected. It is therefore important that PII is handled with care in order to protect the company's brand, maintain the customers' trust as well as to meet the new legislation.

A common pitfall when de-identifying data is to only remove attributes than can obviously be classified as PII, such as VIN numbers. However, as pointed out by Narayanan and Shmatikov [3], defining and identifying PII is surprisingly difficult, and in fact, PII has no technical meaning. A vehicle has approximately 7700 unique signals [4], and while these signals may seem to be separate from PII, even observing a driver's behaviour for as short as 15 minutes is enough to fingerprint and identify a driver with high accuracy [5]. Furthermore, Gao et al. [6] showed that the driving speed in combination with an external road map is enough to trace the location of a vehicle with high accuracy, even though GPS data has been removed. In addition, Tockar [7] demonstrated that an "anonymized" version of NYC cab data, in combination with public data, contained enough information to track celebrities and identify passengers that visited sensitive locations in the city. Thus, all data should be treated as PII, since auxiliary data might be available to re-identify individuals. For example, the position of the car seat might not seem to be PII, but it is likely enough to distinguish between two drivers of the same car.

A promising privacy model with rigorous, mathematical privacy guarantees that could solve the previously mentioned problems is *differential privacy* [8], [9]. Intuitively, for an individual, the best privacy is achieved by not participating in a survey, as their data will not affect any statistics released from such a survey. Consequently, differential privacy aims to approximate one individual not being in the data set. Furthermore, differential privacy's privacy guarantees are robust and does not change over time, as it is *backward and forward proof*. That is, any current or future data set cannot affect the

privacy guarantees offered by differential privacy.

As claimed by Dwork, differential privacy is able to provide high *utility*, accuracy, as well as high privacy in many cases [9]. This is a very desirable property, as there exists a trade-off between privacy and utility that is difficult to balance. Intuitively, this trade-off can be explained by investigating two extreme cases. Without utility, privacy makes little sense, as privacy without utility is satisfied when no data is gathered. However, full utility is achieved by publishing a raw data set, which does not give any privacy guarantees. As neither of these two cases are desirable, a trade-off between the two must be made.

While differential privacy shows promise, it can be challenging to use in real-world cases, as the utility is affected by different parameters. The most prominent real-world cases that use differential privacy have been presented by large companies, such as Apple [10] and Google [11], and only cover very limited use cases. In particular, for vehicular data, differential privacy has so far only been investigated for floating car data (FCD) [12]. Since differential privacy has not yet been established in the automotive domain, although there is a need for privacy-preserving analyses, we believe that differential privacy is a future trend that this paper will aid in paving the way forward for. Hence, the contribution of this position paper is a comprehensible introduction to differential privacy (Section II, III and IV), where we investigate what type of differentially private analyses can be performed in the vehicular domain from a holistic perspective, not only for one specific data type. Furthermore, we provide recommendations (Section V) for how to proceed when implementing differentially private analyses in the vehicle domain, and highlight the challenges (Section VI) involved with the implementation.

II. Differential Privacy

Differential privacy originates from statistical research and examples used often include queries on databases. It is important to note that differential privacy is designed to suit statistical queries that make predictions for large populations, as it prevents inference of information about an entity. As has been pointed out, any meaningful privacy guarantees for differential privacy are not achievable when specific individuals in a data set should be identified [13]. For example, differential privacy will not return any useful information when we ask if Bob uses his company car on weekends.

The differential privacy definition, shown in Definition 1 [9], states that when the same query is run on two neighboring data sets, differing in at most one element, the difference between the probability of getting the same outcome of both queries is essentially negligible. In other words, the presence or absence of one single record does not affect the outcome of a query noticeably. Intuitively, the idea behind differential privacy is to produce a result to a statistical query that is *almost* indistinguishable whether or not one record is present or absent in the data set.

Definition 1 (ϵ-differential privacy): A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,

$$Pr[\mathcal{K}(D_1) \in S] \leq exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S]$$

Two of the main properties of differential privacy are query *composability* and *post-processing* of data [14]. Being composable means that any results of differentially private analyses can be combined, in which case privacy degrades additively. Composability also allows several queries to target the same data. Other privacy models, such as *k*-anonymity [15], fails under composition [16], even with itself. Lastly, any post-processing conducted on data released under differential privacy can be included in any additional analyses, without increased risk to an entity [13].

The risk incurred on an individual is monitored by ϵ, which is sometimes also referred to as the *privacy guarantee*. When ϵ is set to a low value, it gives higher privacy at the cost of reduced utility, whereas a high ϵ gives lower privacy and higher utility. Thus, setting ϵ appropriately is a trade-off between utility and privacy and should be carried out by an expert in the domain.

Another parameter involved is the privacy budget, which is a global parameter from which ϵ is deducted when a query is run. The privacy budget is being consumed by querying the database in order to maintain privacy, and the more queries the higher noise the answers receive. This response can intuitively be explained by an example including the game of twenty questions. In the game of twenty questions, the more questions that are answered, the closer the contestants get to the real answer. To counteract anyone from finding the real answer under differential privacy, the privacy budget enforces that each consecutive answer gets more vague. When the privacy budget is depleted, ϵ can only be set to zero, which means answers will no longer return any meaningful information about the data.

There are many different ways of achieving differential privacy, as any function K that fulfills Definition 1 is differentially private. The reason for why there are many different algorithms is that they are data dependent, and the utility from a differentially private algorithm changes depending on its input data [17]. Consequently, researchers are constantly inventing new algorithms that are optimized for their analysis, resulting in a vast number of differentially private algorithms with varying complexity and utility.

III. Release Mechanisms

The basic idea of a release mechanism, K from Definition 1, is to add probabilistic noise to the real query result. Different release mechanisms are better suited for different data types, such as numerical or categorical data. The lower bound of the accuracy of each release mechanism can also be proven mathematically in order to determine which mechanism is most likely to yield high utility.

Release mechanisms can also be deployed in two different modes: centralized or local. In the centralized mode differential privacy is guaranteed by a trusted party, usually at the time when the database is queried. For local differential privacy on the other hand, each data point is collected under differential privacy in a distributed manner, meaning that noise is added locally. In this section we will describe the Laplace mechanism, the exponential mechanism and randomized response. Figure I shows an overview of the mechanisms and their respective characteristics.

Mechanism Name	Deployment Mode	Answer Data Type
Laplace Mechanism	Centralized (Off-board)	Numerical
Exponential Mechanism	Centralized (Off-board)	Categorical
Randomized Response	Local (On-board)	Categorical

TABLE I: Comparison between the characteristics of three common differentially private mechanisms

A. The Laplace Mechanism

The Laplace mechanism, illustrated in Figure 1, works by adding controlled numerical noise drawn from a Laplace distribution to a query answer. To be able to hide changes in the data set, the query sensitivity, Δf, in combination with the privacy budget, ϵ, is used when generating the noise. The query sensitivity is the maximum impact removing or adding *any* record to the data set has on the query result.

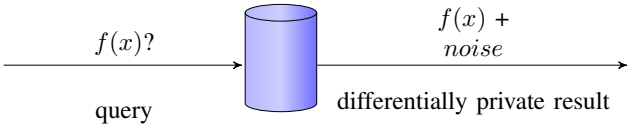

Fig. 1: An illustration of a database with a Laplace mechanism that is used to release differentially private query answers

Since the Laplace mechanism produces continuous numerical noise, it is suitable for queries that are also numerical. Queries can be either continuous or discrete, as differential privacy allows post-processing. In case of a discrete query, the output will be continuous, but can be rounded up to a discrete value without violating differential privacy.

The Laplace mechanism is applied centrally by a trusted party. Thus, all raw data is kept in a database off-board, where each query result is released under differential privacy.

B. Exponential Mechanism

The exponential mechanism [18] is designed for categorical data, so the added noise is not numerical. Rather, the analyst provides a *utility function* that specifies the distance between the different categories. For example, the analyst might want to specify the distance between colors, where shades of the same color are closer than a different color. The exponential mechanism then uses the utility function to output a good answer to the query with higher probability than outputting an answer further from the truth. Thus, the exponential mechanism favors answers that have high utility for a given query input. Like the Laplace mechanism, the exponential mechanism is also applied centrally.

C. Randomized Response

Randomized response [19] was originally invented in 1965 to estimate the amount of people in the population that belong to a sensitive group. Since membership of the group is sensitive, the respondent has an incentive to lie if he or she is part of the group, which can cause a skewed distribution of answers. Therefore, randomized response provides a protocol which gives the respondents *plausible deniability*, meaning that an analyst cannot tell if a given respondent lied or not while still being able to make predictions about the population.

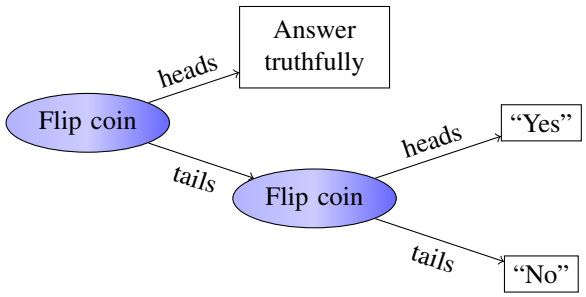

Fig. 2: Randomized response, in this example following the protocol to answer the question "Do you text and drive?"

Randomized response enforces local differential privacy, and each driver follows the protocol in Figure 2 in order to respond under differential privacy. In order to interpret the results from randomized response, the analyst has to extract the number of people that where telling the truth using Bayes' theorem.

IV. Privacy Guarantees

In order to utilize the privacy budget well, making it last longer than when using a naïve approach, privacy can be applied at event-level [20] rather than user-level. Event-level privacy protects a single event, such as a single data point where a driver is speeding, whereas user-level privacy typically protects an individual or an object such as a vehicle. The analyst defines what an event is, for example a reading of one single signal or something that happens after a certain condition is met. For example, one event might be that the airbag has been triggered, but it could also be one single reading of the engine temperature.

Essentially, the privacy level determines what or who should be protected by differential privacy, by determining what data points are considered to belong to one entity. In other words, if we choose user-level privacy for a car, all 7700 signals belong to that entity, whereas if we decide on event-level privacy, we can decide on a subset of those signals.

V. Advice

In theory, any query can be answered under differential privacy. In practice, however, some queries are better suited, since they offer a better trade-off between privacy and utility. Hence, in this section we will present some advice regarding how to proceed when creating a differentially private analysis for vehicular data.

A. Model the Domain

1) Decide the privacy level: Before starting to implement anything, it is important to define who or what privacy should be provided for. For example, if the driver's identity should be protected, user-level privacy needs to be used. Also, since a driver can drive more than one vehicle, this needs to be accounted for in the model.

In some cases, to improve the utility of the answer, the analyst might settle for only hiding certain events, such as speeding, in which case the analyst can choose to only provide privacy for the speed of the car. On the other hand, the analyst can also choose to hide only the time a driver was driving at a certain speed. In the case where only the time is hidden, the driver can deny that he or she was speeding since it is impossible to infer where the driver was driving. In other words, an analyst can choose to hide events of different sizes, such as only the time something happened or an entire driving pattern, and it is vital to define in advance what those events are.

Thus, modeling the kind of privacy that should be given and to whom needs to be done first, in order to decide the privacy level as well as finding a suitable value for ϵ.

B. Trusted Party or Not?

1) Decide deployment mode: The main advantage of local differential privacy is that each driver adds their own noise, as opposed to centralized differential privacy. Thus, local differential privacy, which can be implemented using randomized response, is carried out on-board whereas centralized differential privacy must be implemented off-board. Since randomized response is local, no trusted party is needed to gather all data, which also means companies never have to store or even get in contact with any sensitive data as it will be kept in the vehicle. Furthermore, on-board algorithms can also result in data minimization, meaning that less data is gathered from the driver, which is a property that is being promoted by the upcoming GDPR. However, the downside of local mechanisms is that achieving an adequate trade-off between privacy and utility is difficult in real-world cases [21].

C. Using the Privacy Budget

In order to get a good balance between utility and privacy, the privacy budget needs to be used with care. We believe there are certain techniques that could make the budget last longer, such as personalized budgets [22] (as opposed to a global budget) and random sampling.

1) Personalized budgets: First, personalized budgets for differential privacy allows each record to keep its own budget, which means all records are not affected by queries that do not concern them. Using personalized budgets thus allows an analyst to keep the budget from being spent unnecessary, as he or she can query all vehicles of a certain model without also spending the budget for vehicles of other models.

From a data management perspective, another benefit of using personalized budgets is that even if there is no centrally controlled database gathering all the data, deductions to a global budget do not have to be communicated across databases as long as all data belonging to one entity remains in one database. Thus, a company can still keep several databases for different kinds of data without introducing dependencies between the databases.

2) Random sampling: Secondly, random sampling allows us to select a subset of records to query, and thus together with personalized budgets only spend the budget of that subset. Random sampling is especially appealing for big data sets, where a subset of the entire population still gives a good prediction. We believe that vehicular data fits this description.

3) Streaming data: Furthermore, we also believe the vehicle industry could benefit from enforcing differential privacy on streaming data instead of storing raw data in an off-board database, as all stored data would be sanitized. That is, vehicles could be selected to be part of a query, and then their replies could be released under differential privacy where the data is aggregated. In this way only the results from differentially private queries could be saved, and raw data thrown away. Since differential privacy offers post-processing, the data kept could then be used in any analysis. Apart from preserving privacy, this approach could also save storage space on the server side, and could also decrease the traffic used to upload data when queries only are issued on demand.

In the case of the streaming paradigm where vehicles are queried, each vehicle would have to keep track of its own budget and communicate it to the server, which would be possible when we use personalized budgets. Even though local differential privacy inherently is better suited for this setting, we believe this provides an alternative where local algorithms offer low utility.

D. Population Statistics, Never Individual Data

Differential privacy is designed to answer statistical queries that make predictions about the population, not for inferring information about individuals. Thus, if an analyst were to ask how often Bob uses the parking brake per week, the result would not be useful as the noise would likely be too high.

The accuracy of results can be vital if safety-critical functionality is to be developed from an analysis. In such cases, the upper-bound and lower-bound accuracy of a differentially private algorithm needs to be calculated before the analysis is carried out. If the differentially private algorithm does not provide a tight upper- and lower-bound on accuracy, the safety-critical functionality could be at risk by using data under differential privacy.

In these cases, there are two options: either the differentially private algorithm is modified (for example by rephrasing the query, see Section V-E) to achieve higher accuracy, or the analysis is carried out without guaranteeing differential privacy on the company's own vehicles. For example, a case where differential privacy is not suitable is for function testing using high-resolution data from few vehicles.

E. Rephrase Queries

Rephrasing a query might result in better utility.

1) Target the population: In some cases an inappropriate query, that targets individuals, can be rephrased into a query that targets the entire population. For example, if we want to find out when an engine is running outside of its specification, asking for in which vehicles this occurs would be a bad idea. On the other hand, what we are really interested in might not be which those cars are, but rather how many they are, to determine if it is common or not. In such a case it is possible to turn a bad query into a prediction about the population, a counting query in this case, which would provide a better answer to, approximately, the original query.

2) Change the query type: In other cases, the problem might not be that one individual is targeted, but that the query itself is prone to result in high noise. As an example, instead of asking for the average speed, the speed can be investigated from a histogram from which heavy-hitters can be identified. In other words, when the query sensitivity is high, transforming the query into a less noisy one is advisable, unless the difference between the query result and the proportional noise is small.

F. Dealing with Query Sensitivity

One issue with query sensitivity is that in practice it can be hard to define. Therefore, in some cases, the query sensitivity needs to be set to the physical maximum of a parameter, which is unlikely but necessary.

1) Query a large data set: Some queries, such as sums and averages, tend to have high query sensitivity. For vehicles, the analyst might then when defining the query sensitivity refer to the maximum value that can be held in a certain register in the vehicle. While these queries can still be used, the noise will be easier to hide when a larger data set is queried. Thus, the data set's size is more important in cases where the query sensitivity is high rather than in cases where it is constant, such as counting queries and histograms.

2) Fixed sensitivity through cropped ranges: The way we suggest for dealing with high query sensitivity is to crop the ranges and set a fixed max and min value. All values outside of range must not be used in the analysis, as they would not be protected in this case. The chosen range itself also leaks information about what range is expected to be normal. When the range itself is sensitive data, the range must be decided under differential privacy.

However, if the range is not well-known, it is possible to accidentally set the range to an interval which a large part of the values fall outside of. To be able tweak an incorrectly set range in a differentially private manner, we suggest creating one bin on each side of the range that catches all outside values. When the side-bins are fuller than a certain threshold, it indicates a problem with the chosen range, which then needs to be redefined.

G. Applicable Analyses

1) Histograms and counting queries: Histograms and counting queries are particularly suited for the Laplace mechanism, as pointed out by Dwork [23]. The reason for this is that histograms and counting queries have a fixed sensitivity, which generally results in low noise that is independent of the data set's size. Consequently, when the data set queried is small, histogram and counting queries are especially appropriate.

2) Numerical queries: Any other numerical query is also possible to implement under differential privacy using the Laplace mechanism. However, the Laplace mechanism is highly dependent on the type of query being asked, as each query type has its own sensitivity, Δf. For example, if we want to calculate the average speed of a vehicle, we need to account for the largest possible change adding or removing any data point to the set can have on the average. Consequently, we must assume the worst case, which in this case is adding the highest possible speed to the data set. Thus, the sensitivity is the difference between the maximum and minimum speed *possible*. The sensitivity will then affect the proportion of noise that is added to the query result, and thus we suggest choosing a query which has lower sensitivity as it generally will yield lower noise than a high sensitivity query.

3) Categorical queries: For data where adding noise makes little sense, such as categorical data, the exponential mechanism can be used. One such example is when asking for the most popular car colors, as adding numerical noise to colors does not make sense. Another example would be if we want to find out what button on the dashboard is pushed the most times.

VI. Challenges

There are many challenges with properly implementing a differentially private analysis in real-world cases. In this section we point out some of the most prominent ones for vehicular data.

A. Setting the Privacy Budget

To reason about ϵ, the domain must first be modeled in such a way that the entity to protect has been defined through setting the privacy level. ϵ is then the factor of indistinguishability between any two entities. Consequently, setting ϵ to a meaningful value is difficult, as ϵ is a relative measure of privacy risk [24]. In other words, the appropriate value of ϵ is affected by the type of data being released. Thus, the risk of two differentially private algorithms cannot be compared by their value of ϵ. This problem is not unique to vehicular data, but follows inherently from the definition of differential privacy.

While how to choose ϵ appropriately remains an open research question, Lee and Clifton as well as Hsu et al. propose practical solutions to the problem. Lee and Clifton suggests choosing ϵ based on the individual's risk of being re-identified [24], whereas Hsu et al. [25] propose that ϵ should be chosen based on an economic model. While no approach is clearly better than the other, both solutions provide an interpretation of what the privacy guarantees mean to a participant, making it possible to communicate the risk accordingly.

B. Multidimensional Time Series Data

Compared to other systems, preserving the privacy of vehicles is particularly difficult since they are highly complex systems that generates vast amounts of data from thousands of signals. To make matters worse, vehicle signals can be gathered continuously over time. Consequently, as the amount of available data simplifies identifying a particular vehicle, hiding the presence of a specific vehicle in the data set becomes more difficult than hiding fewer connected data points.

Because of the multidimensional time series nature of the data, performing more than one analysis with high utility that guarantees user-level privacy becomes infeasible. User-level privacy would also not allow the analyst to reset the budget, not even after years of using the same budget. Consequently, we believe that in order to maintain utility, analyses can only provide event-level privacy.

On a positive note, providing event-level privacy can save the manufacturer the trouble of maintaining the privacy budget between different systems, as it results in separate privacy budgets for each system.

An open issue that we need to solve in this area is interpreting what event-level differential privacy means for a driver, as it is an individual that ultimately wants the privacy. For example, what does it mean from a privacy perspective if we only hide at what point in time the battery had a certain temperature? Event-level privacy might be more feasible than user-level privacy from a utility perspective, but every case must be investigated to make sure the privacy guarantees in such a situation makes sense to an individual as well.

VII. Conclusion

For vehicular data, differential privacy can be especially tricky to enforce due to the fact that vehicles contain a system of thousands of dependent signals collected over time. Consequently, the automotive domain is very complex from a privacy perspective. However, as differential privacy is the only privacy model that provides provable privacy guarantees, this is currently the only robust way of mitigating re-identification attacks on data while maintaining utility. Thus, we believe that the automotive industry will benefit from carrying out their privacy-preserving analyses under differential privacy.

In order to properly implement differential privacy, it is vital that the company first model the privacy within their domain, to determine what they are trying to protect. From the model, the company can then define what signals an event should consist of, and the model also makes it possible to reason about a suitable value for ϵ. Only after the modeling has been done can the implementation details of the analysis be decided.

Differential privacy should be used to answer statistical questions about a population. Since differential privacy aims to protect the privacy of each entity, it is not suitable for detecting anomalies. Because of this, analyses on high-resolution data from few vehicles, such as when performing function testing, should not be carried out under differential privacy. Any other statistical queries can be answered under differential privacy, but we believe that one of the main problems with introducing differential privacy in the automotive domain is maintaining high utility for the analyses. Thus, we have investigated ways of being able to spend the privacy budget wisely.

We believe that in order to enforce differential privacy for vehicular data in a sustainable way, personalized budgets, random sampling as well as event-level privacy are key to high utility. Rephrasing queries as well as cropping ranges of queries is also something that can make differential privacy more applicable. Furthermore, we believe that by issuing queries to vehicles on the go using the streaming paradigm or local differential privacy, there is potential to save both storage space and bandwidth while preserving privacy at the same time.

In the end, we believe differential privacy shows promise for the vehicle industry. However, more work still needs to be put into interpreting the meaning of ϵ as well as event-level privacy from a customer's perspective, as the meaning will differ on a case-by-case basis.

References

[1] FEDERATION INTERNATIONALE DE L'AUTOMOBILE (FIA) REGION I. What europeans think about connected cars. (visited on 2017-01-24). [Online]. Available: http://www.mycarmydata.eu/wp-content/themes/shalashaska/assets/docs/FIA_survey_2016.pdf

[2] European Parliament, Council of the European Union. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). (visited on 2016-09-06). [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&qid=1473158599287&from=EN

[3] A. Narayanan and V. Shmatikov, "Myths and fallacies of "personally identifiable information"," *Commun. ACM*, vol. 53, no. 6, pp. 24–26, 2010.

[4] P. Kleberger, N. Nowdehi, and T. Olovsson, "Towards designing secure in-vehicle network architectures using community detection algorithms," in *2014 IEEE Vehicular Networking Conference (VNC)*, 2014, pp. 69–76.

[5] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 1, pp. 34–50, 2015.

[6] X. Gao, B. Firner, S. Sugrim, V. Kaiser-Pendergrast, Y. Yang, and J. Lindqvist, "Elastic pathing: Your speed is enough to track you," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14. ACM, 2014, pp. 975–986.

[7] A. Tockar. Riding with the stars: Passenger privacy in the NYC taxicab dataset. (visited on 2017-02-15). [Online]. Available: https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, ser. Lecture Notes in Computer Science, S. Halevi and T. Rabin, Eds. Springer Berlin Heidelberg, 2006, no. 3876, pp. 265–284.

[9] C. Dwork, "Differential privacy," in *Automata, languages and programming*. Springer, 2006, pp. 1–12.

[10] A. Greenberg. Apple's 'differential privacy' is about collecting your data — but not your data. (visited on 2017-01-30). [Online]. Available: https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/

[11] U. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14. ACM, 2014, pp. 1054–1067.

[12] F. Kargl, A. Friedman, and R. Boreli, "Differential privacy in intelligent transportation systems," in *Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks*, ser. WiSec '13. ACM, 2013, pp. 107–112.

[13] S. Vadhan, "The complexity of differential privacy," (visited on 2017-02-06). [Online]. Available: http://privacytools.seas.harvard.edu/files/privacytools/files/complexity privacy_1.pdf

[14] C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! a survey of attacks on private data," *Annual Review of Statistics and Its Application*, vol. 4, no. 1, 2017.

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

[16] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 265–273.

[17] M. Hay, A. Machanavajjhala, G. Miklau, Y. Chen, and D. Zhang, "Principled evaluation of differentially private algorithms using DPBench," in *Proceedings of the 2016 International Conference on Management of Data*, ser. SIGMOD '16. ACM, 2016, pp. 139–154.

[18] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science, 2007. FOCS '07*, 2007, pp. 94–103.

[19] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[20] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, "Differential privacy under continual observation," in *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, 2010, pp. 715–724.

[21] R. Chen, H. Li, A. K. Qin, S. P. Kasiviswanathan, and H. Jin, "Private spatial data aggregation in the local setting," in *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*. IEEE, 2016, pp. 289–300.

[22] H. Ebadi, D. Sands, and G. Schneider, "Differential privacy: Now it's getting personal," in *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, no. POPL'15. ACM Press, 2015, pp. 69–81.

[23] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, ser. Lecture Notes in Computer Science, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Springer Berlin Heidelberg, 2008, no. 4978, pp. 1–19.

[24] J. Lee and C. Clifton, "How much is enough? choosing ϵ for differential privacy," in *Information Security*, ser. Lecture Notes in Computer Science, X. Lai, J. Zhou, and H. Li, Eds. Springer Berlin Heidelberg, 2011, no. 7001, pp. 325–340.

[25] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *2014 IEEE 27th Computer Security Foundations Symposium*, 2014, pp. 398–410.