Privacy-Preserving Data Collection Frameworks for Autonomous Vehicle Telemetry

By Dr. Natalia Borisova

Associate Professor of Artificial Intelligence, ITMO University, Russia

1. Introduction

Despite clear demand for this kind of real-world experience data, it is difficult to see how to sustainably and widely collect and use it. The recent debates about the role of third-party data and direct vehicle-to-government data reporting in new federal guidance for vehicle safety compliance and testing provide a leading-edge illustration of just how expensive and contentious it might be to create any kind of centralized collection infrastructure for this kind of data. In this paper, we present several privacy-preserving data collection frameworks that collect the semantics of physical-world events and aggregate the probe data inside the autonomous vehicle before transmitting them in raw form, after appropriate filtering and reduction, to a cloud-based data broker. We discuss the tradeoffs and limitations of this approach. Our protocols contribute to the formation of a marketplace to inform law enforcement, regulators, insurers, and the public about the availability, format, and prices of different types of data that are desired to produce the kind of transparency and accountability that society will expect from those that will be allowed to purchase, build, operate, and benefit from autonomous vehicle technology.

Autonomous vehicle technology holds great promise to revolutionize personal and commercial transportation and make mobility more convenient, efficient, and accessible. This promise, however, is accompanied by a host of challenges with respect to public safety, consumer trust, integration with legacy modes, insurance and liability, and cybersecurity, to name just a few. A recurring theme across many such challenges is the importance of robust, empirically-based evidence about how autonomous vehicles actually perform in the real world in order to enable law enforcement, regulators, insurers, and consumers to trust that vehicles will work as intended in a wide variety of conditions. For example, in the unfortunate event of a crash or other property damage, incident investigators will seek to understand whether the vehicle was functioning properly at the time of the event, and incident investigators, law enforcement, and autonomous vehicle manufacturers may benefit from real-world data about incidents to understand the nature of public risk and to shape responses.

1.1. Background and Motivation

There are many reasons to be concerned about the security and privacy of the already big and the potentially huge amount of data generated and used in several phases of an AV, such as design, implementation, and utilization. Its volume might grow to dimensions, which make it comparable to Big Data and might lead to a massive increase in private surveillance. Furthermore, this is federated data because it belongs not only to the different owners of the individual AVs, but also to other entities, among them the manufacturers and, especially, data providers, such as designers, engineers, and technicians. If these data are used with specific goals, the owners of the associated companies are not only the AV owners and of the car manufacturers, but also of the telecommunication and data storage companies involved in data processing for safety-critical applications. This situation is reminiscent of the Critical Infrastructures (CI) which enable energy, water, finance, transportation, government, defense industrial base, and communications.

The current trend in both the automotive sector research and the industry for Autonomous Vehicles (AVs) is to deploy them in the near future. For this, improving very diverse and numerous aspects is a priority. Among them, one of the most outstanding is the need to guarantee the safety and well-being of the passengers, which involves a suitable management of the data generated by the Vehicle Telemetry (VT) hardware/software equipment, the corresponding communications hardware/software, and the commanded manipulators. Recent reports have shown that AVs collect and process, often off-device, (geolocated or semigeolocated) telemetry data of the vehicle's environment and its passengers and communicate with extremely diverse actors, some of whom are external to the AV and its passengers.

1.2. Research Objectives and Scope

The key topics of the study include the privacy issues of California's AVs and framework laws, what data can be collected from us without consent, the identity of the public in time series data, the perspective of the residents of California, and what and why we should manage privacy matters. The paper proposes a forcing technology framework. The principals and suggested courses of realization in the driving system have a clear relationship with the law prohibiting the collection, use, recording, and recording of driver and user data without consent and use this framework in most cases to protect driver privacy. These concepts and objectives are extended and introduced to the field of collecting AV data of people on public roads and are consistent in the same matter as the concept of the driver and user. This will help you understand and make clear the driver's problem of capturing AV data in public.

Proudian, Grusche, Keeney, Kreid, and Moses of UC Berkeley performed a self-driving system test in 189 miles around the city of Los Angeles. That test collected video data and dozens of other sensor data from local landmark road sections. The test's algorithm identifies and separates individuals in public photos and videos from multiple autonomous vehicle sensors in real-time. By carefully selecting the location and testing for the data collection and algorithm, the autonomous framework system meets all California laws and regulations and protects the collection of individual privacy data.

- To apply the privacy-saving data policies to real implementation and discuss their effectiveness in real data collected in various busiest street districts in the Los Angeles area.

- To establish and implement necessary company policies in data handling, security, and privacy policies to make the technology process responsible and transparent.

- To formulate practical technical suggestions on the balance of advanced technologies with the laws and regulations for handling privacy matters, while ensuring safety, autonomous operation, and data protection under the strict requirements of the California Public Utilities Commission, the Personal Information Protection and Electronic Documents Act of Canada, and the General Data Protection Regulations (GDPR) of the European Union.

The objectives of this research for the privacy-saving nature are listed below:

2. Autonomous Vehicles and Telemetry

To test and train the control logic of autonomous vehicles, vehicle developers utilize trial cars to collect telemetry data during road testing in different locations, conditions, and situations. Vehicle development companies require a lot of proprietary vehicle telemetry datasets to make profitable business decisions. Vehicle manufacturers sometimes consider adding specialized telemetry collectors to mass-produced cars because such collectors can help perform market research and expand the geographic and demographic representation of available data. Because of the need for extremely large diverse datasets, vehicle development companies might create an incentive ecosystem to compensate vehicle owners who help by sharing direct access to their real-world travel experiences. Vehicle owners can sometimes obtain additional benefits if the vehicle telemetry data is stored in a private data storage that requires permission to change its onboard data settings, e.g., by signing with a private key. The influx of vehicle personal information obtained and stored by vehicle manufacturers, however, raises more robust privacy preservation requirements. A public-private partnership that relies on secure and private data storage can be used to enforce transparency and accountability.

The growing popularity of autonomous vehicles is transforming the driving experience. Substantial private and public support is backing a wide array of autonomous vehicle development efforts. Autonomous vehicle technology aims to minimize accidents, optimize road capacity, improve traffic flow, eliminate the need for parking space, reduce energy consumption, cut down on idle times, and provide lower transportation costs. The largest barriers to unlocking the potential of autonomous vehicles are technology advances and cost. The autonomous vehicle development process typically requires the collection and analysis of a large volume of data to develop safe control logic, which poses ethical and privacy challenges that need to be addressed.

2.1. Overview of Autonomous Vehicles

The motion control sub-system operates based on the decision-making algorithms, maps, environment models, and predictions. The localization system identifies the vehicle's position in the current map with respect to location and pose (3D position and orientation); as a result of pose estimation, the AV has inferred a new location/pose and updated its location data. The motion prediction system is dedicated to the prediction of the expected movements of other agents in the environment; for passenger safety, the autonomous vehicle must also predict the behavior of pedestrians, cyclists, or other vehicles that are not yet in the vehicle's proximity but will be present in it after a few time steps. The high-level behavior generation is an AV system function that meets high-level objectives, such as monitoring a desired path and driving in a certain lane; the synthetic evolution of each of these behaviors will be context-driven and must align with the AV's overall objectives.

An autonomous vehicle (AV) is described by its treatment of information flow from input to output. The combination of sensors for the AV vehicle carries some sensing tools such as cameras, radar (Radio Detection and Ranging), LiDAR (Light Detection and Ranging), and IMU (Inertial Measurement Unit) for gathering information about the specific manner in which the car perceives the environment. The motion control refers to the individual and combined actions that the AV takes to affect its own and other entities in its surroundings. Examples of motion implementations include actuating individual wheels or controlling brake or steering in the vehicle, and signaling a turn or change of direction.

2.2. Importance of Telemetry Data

As an example that demonstrates the need for accurate telemetry data originating from vehicles, car rental companies as well as Uber drivers endeavor to calculate fees based upon miles driven and road use. Taxi services can use mileage and location data to monitor their vehicle and driver usage, ensuring that cabs are not used in excess of their useful life, suppressing drive and road wear when it would be socially inefficient to do so. Emergency road services dispatchers would also benefit from knowing more about car faults as well as a car's particular technical configuration such that repair estimates and road capacity calculations could be improved. Various service providers could benefit from knowing how often their customers visit a given region or how long the customer spends traveling versus engaging in other useful activities. Goods delivery services want to know not only the most efficient set of paths to take but also how emissions are related to miles driven. Companies using fleets of vehicles, for instance, to deliver goods or provide services will benefit from automated and ongoing monitoring of vehicle health, making it possible to maintain vehicles before problems escalate. Finally, the scientific community is also increasingly interested in the study of emissions and patterns of use as a base to validate climate and abatement models. The research of Teubal et al. meets some of the car rental company's practical and ethical demands of careful data collection, and since embedded loggers are expensive, sharing this type of data is the basis for economic argument as well as goodwill to be developed between smart car companies and owners alike.

The evidence for the importance, benefits, and usefulness of collecting and analyzing telemetry data from automobile systems is abundant in existing literature. Automakers and other firms are keen on understanding how cars perform in the wild; even cars of the same make and model can have very different lifetimes and failure rates depending on the actual use patterns of the cars and road conditions in the specific geographical regions the cars are driven. In addition, automobile firms are typically interested in understanding how drivers use feature sets of interest to them such as car networking features, infotainment features, and advanced driver assistance systems. As connected cars become increasingly common, companies investing in providing the infrastructure that makes connected cars possible are interested in understanding, for a wide range of real-world drives, the resource consumption and quality of service of car networking as it pertains to driver and passenger use and experience. Those companies are also interested in determining the environmental impact of shared cars through the collection of real-time feedback on emissions.

3. Privacy Concerns in Data Collection

As a prior step, this study identifies the privacy risks during various stages of collecting vehicle telemetry data and provides a data privacy framework. The proposed framework aims for the collection of usage-dependent vehicle telemetry data, which is minimally required to be processed in a vehicle-area network to address the data privacy concerns, specifically for a future autonomous ride-sharing fleet of electric vehicles. To build the framework, privacy risks were analyzed by applying engineering methodologies, such as trust boundaries and Functional Failure Modes and Effect Analysis (FFMEA), and utilizing domain-specific knowledge from persons associated with related risk analyses and from experts with legal and data knowledge. For validation, the framework was developed by mapping numerous vehicular data points from actual autonomous vehicle data to understand the data from the perspective of privacy legislation. The concerns and the framework are discussed further with datasets from connected vehicles.

Many valuable insights can be learned by collecting large amounts of diverse data generated by autonomous and connected vehicles. However, privacy has emerged as a significant concern in vehicle-related data collection and analysis. Visual data, which can be captured through cameras in connected and autonomous vehicles, can include a significant amount of personal information about the vehicle's driver and possibly other pedestrians. Hence, legislation is being adopted or is under consideration by various bodies to require minimization, encryption, pseudonymization, or even the destruction of personal data collected by vehicles in the context of the driving task. For example, the European General

Data Protection Regulation (GDPR) has a significant impact on how personal data is handled in the context of driving, stating that the data might be processed only for a specific purpose and within legal constraints, and that data processing inside the vehicle should be minimized. Similarly, the United Nations Global Cyber Regulation for Traffic Safety also includes considerations for privacy-preserving external communications of vehicles.

3.1. Types of Data Collected

According to the function, data type, and level of abstraction of the data collected by an AV, the data can be further classified as geolocation data, kinematic data, and non-kinematic data. Geolocation data are telemetries that contain geographic location and location precision, but without additional data like time or data type collected by an AV. The collected data with no time and location are used to diagnose sensors and actuators inside the data. Kinematic data are telemetries that are collected from the kinematic controller. This usually captures the vehicle's instantaneous lateral and longitudinal motion, such as heading angle, yaw rate, longitudinal acceleration, longitudinal velocity, and lateral acceleration of an AV. Non-kinematic data are telemetry data that do not belong to the kinematic group. Depending on the function of the collected data, kinematic data can either be telemetries collected by an automated driving system at high frequency to provide updated vehicle state propagation information, collision detection/judgment/control, or telemetries collected at regular low frequency intervals. The latter contains aggregated trajectory details, vehicle behavior information, or other AV control/traffic related information. These are provided to guide the network optimization/sensitivity analysis/modeling process in the next place with minimum potential data privacy risk.

The usage of autonomous vehicles (AVs) for urban transportation is expanding rapidly. The data collection related to AVs falls into two categories: passive and active collection. In passive collection, an AV collects the data independently based on its usage, environment, and state-of-the-art hardware and software. It uses the collected data to perform its initially intended functions. In active collection, multiple agents, such as transportation authorities and traffic management agencies, guide the way to collect the data. This is done to address challenges like urban mobility, traffic flow management, and road safety. The collected AV data usually contains information including speed, acceleration, and in some cases, the status of sensors and actuators in an AV. This data is also called telemetry.

3.2. Risks and Implications

Such threats are very realistic and should not be taken lightly, as they are highly demotivating for data production and sharing, undermining the goals of producing high-quality dataset. To address these emergent privacy concerns, various privacy-preserving data collection frameworks are currently being developed, which aim to balance the trade-off between data sharing and the individual's data privacy protection and sensitivity.

Due to the nature of privacy-preserving the involved data collection operation might pose vehicle telemetry, there are severe implications and adverse outcomes if not enough precautions and action remedies are taken. Several inadvertent disclosures can easily be exploited by cybercrime organizations or individual black hat hackers, including the physical location of the vehicle and the owner's point of interests. These can be utilized to perform a wide range of attack vectors, such as burglary and theft, especially because the malicious entities can exploit vulnerabilities discovered while both the vehicle and/or the vehicle owner is away. Furthermore, a deep analysis of the owners' driving patterns could expose sensitive personal information, i.e. work address, work hours and routes, which could lead to exposure of vulnerabilities, such as exploitation on the road, kidnapping or other that could jeopardize the owner's safety and well-being.

4. Existing Privacy-Preserving Techniques

A. K-anonymity K-anonymity is a simple and efficient privacy-preserving technique. The basic principle is the following: in each collection of data items, all the data items pertaining to any one individual should be indistinguishable from a data item pertaining to at least k-1 other individuals. K-anonymity has been used in a broad range of areas. In location privacy, location cloaking mechanisms provide k-anonymity - all location updates produced by the same group of individuals (up to a maximum distance w) are approximated to be at the geographic center of the group. In webcam systems, both the original and the modified (faces replaced with public-domain cartoons) webcam images must satisfy a form of k-anonymity based on the background of the images. K-anonymity can be generalized in two directions: l-diversity (or (k,l)-anonymity), where each equivalence class should contain at least l sensitive value instances; t-closeness, where the distribution of the values of the sensitive attribute should be close to the distribution of the values of the sensitive attribute in overall data.

In this section, we provide an overview of the existing privacy-preserving technologies that are suitable for the proposed framework.

4.1. Anonymization and Pseudonymization

Recognizing the non-trivial, non-reversible, and non-assuming properties required in real-life applications, privacy-enhancing techniques that are acknowledged to be sufficient, such as k-anonymity (surname identifying threshold k), are negatively shaped. In contrast, differential privacy has been proposed as a state-of-the-art privacy definition that lies in the fact that the output of any analysis shall not reveal whether any individual is present in the input data. The differential privacy framework can well safeguard the individuals in a population against arbitrary and unrestricted access to information, no matter at which point in time the queries have been executed. Differential privacy is declared to provide a strong and quantifiable privacy guarantee and is pursued for the challenging problem in security and privacy research and for releasing telemetry data.

In contrast to anonymization, in pseudonymization, individual IDs are replaced with non-identifying and distinct numerical values or keys (pseudonyms) to make the data non-trivially disparate to be associated with any known individuals. As signaled by the terms of the GDPR, such as sensitive data and personal data breaches, both anonymization and pseudonymization can be used to guard intruders against the key information with regard to the individuals. Hence, they can be used to safeguard the person's identity.

Anonymization and pseudonymization have been widely adopted as privacy-enhancing techniques in security and privacy research. The methods use different approaches to delink an individual's identity from their data. Anonymization aims to remove the person or key attributes in the primary data so that the data can no longer be associated with a particular individual. This is generally achieved by removing an individual's direct identification (ID), such as their name or phone number, from the data.

4.2. Differential Privacy

Utilizing differential privacy in machine learning training includes adding noise to individual gradients with its inherent security implications. However, with such a mathematical guarantee, this becomes a safe protection measure for protecting individuals' privacy. Differential privacy has been defined in the context of machine learning and released models that the "model's outputs will not just be associated with details (datapoints) of the training data but can be proven to be independent of each individual training data with orders of magnitude difference". Notably, this privacy definition considers both single vs. auxiliary data and does not include the arguments of training iteration. This makes differential privacy a more attractive privacy-preserving data release framework for machine-learning based systems.

Differential privacy is a robust privacy protection framework that provides personal information privacy to people whose data is used to drive digital services. Consider a scenario where a question concerns a set of house items, and the answer to the question should not disclose whether S (e.g., with cancer) belongs to that set. An algorithm satisfies "#-differential privacy" if, for all possible answers to a question, the probability of any particular answer differs from the probability of any other answer by a factor of at most #, for privacy level #. Individual differential privacy provides (' ε = 1)' differential privacy guarantees - a small enough value to ensure that a query response will be effectively randomized. In that particular setting of differential privacy, we have a guarantee that if an individual (person's data) is included in the dataset of the algorithm that answers the question, the probability of the specific output being generated by that individual's inclusion is restricted to a small value (1/e) vs. the scenario when the data of the individual is not included in the dataset.

5. Challenges and Limitations

However, after the large-scene segmentation, data is still in large volume. Existing secure multi-party computation algorithms require a large volume of communication. An efficient and secure approach of generating data partitions in the agreement of all respective partitions among different participants is imperative to good data transfer methods SAFER-KNN.

5.1 Scalability An autonomous vehicle can produce large quantities of data per hour. While for experiments we only had the opportunity to collect several hours of data, scaling up our work to larger experimental datasets has not been a significant obstacle, despite the fact that adult individuals can already achieve high ACC using their different driver recognition models. This is because any large-scene segmentation or clustering can be efficiently and securely parallelized.

The framework still needs further study and development. In this section, we discuss some initial thoughts on the challenges and limitations of the proposed approach.

5.1. Accuracy vs. Privacy Trade-offs

Reducing the number of times the infrastructure receives episodic high-frequency raw telemetry data with the lowest level of abstraction can help mitigate the risk of privacy breaching, but unfortunately, it could come at the cost of richness of the training dataset. By converting the raw data into a more abstract higher-level data and then using an anonymizing protocol instead before uploading it to the cloud, the proposed architecture allows this accuracy vs. privacy trade-off to be a knob that can be tuned. The anonymizing protocol can be a computationally-intensive non-interactive privacy-preserving computation. The participants in this computation are data owners that the anonymized raw data have label governance enforced. Control can be given in certain use-cases such that a data owner with a level of trust (e.g. consortium members, within the enterprise) is not subject to a certain expensive anonymizing protocol, but others are.

In the case of a non-privacy-aware scenario, with the rich number of sensors in autonomous vehicles, a vast amount of telemetry data could be streamed out to the cloud. The first phase of a generic machine learning pipeline is training, which can be accomplished using labeled telemetry data collected from the fleet. The more natural cluster many driving maneuvers belong to, the better the driving behavior models could be trained. Serialization of these models and pushing them onto the fleet of vehicles for on-device inferencing is a well-discussed strategy in the literature given the computational budget and constraints. This is the second phase of the machine learning pipeline and typically the runtime in which models are run against telemetry to derive insights on the state and usage of these vehicles.

5.2. Scalability Issues

On one hand, a smaller batch size would lead to a high communication overhead and long hardware utilization of iDoc; On the other hand, a larger batch size exhibits the consequent privacy-risk behaviors, which may reveal more driving details to the cloud side. Consequently, the intelligence inspector recommends selecting a determined performance trade-off, such as six intervals within the originally defined interval to obtain the optimal batching size.

To overcome this problem, in the proposed variant Flush-Histogram, each AV only contributes its local histogram when the local observed speed interval range number of increments equal to batching size. Specifically, batch size denotes the number of incremental wheeling degree histograms along the speed-up–bin axis.

The original HEGA scheme is incremental in the sense that the edge server should open the encrypted histogram from the wheeling degree counter at each EI based on incrementally obtaining the corresponding encrypted wheeling degree data with common speed ranges from different associated AVs. Although we can synthesize an immediate histogram sharing mechanism to speed up and save the bandwidth, this is still higher communication overhead compared to the previous incremental ECC schemes.

On one hand, a smaller determined step means a longer communication interval and regular prompt response times for AVs. On the other hand, a larger determined step implies shorter communication intervals and faster real-time security scores for AVs.

Given a large number of connected AVs and a continuous observed speed range, the global observed speed range would be expressed as a sequence of non-overlapping local observed speed interval ranges at a certain resolution (we refer to this resolution as the determined step), which would be arbitrarily partitioned during different evaluation interactions (EIs) for all associated AVs.

In the original HEGA scheme, the AV specifies the range of their local observed speeds to the edge server during the crypto-computation. At each round, the edge server can obtain corresponding encrypted histograms from all associated AVs within this range. Due to the maximal permissible tolerable communication delays, the observed speed interval range should be inherently determined by the AV before the next submission.

In this section, we discuss several aspects related to the scalability of HEGA with respect to both the communication load between the AVs and the edge server, and the computational complexity of the edge server at the cloud side.

6. Proposed Framework for Privacy-Preserving Data Collection

In this paper, we focus on UAV-assisted data collection to address the privacy-preserving issues in collecting autonomous vehicle telemetry. We investigate an integrated telemetry acquisition and data mining scheme leveraging mobile edge computing on unmanned aerial vehicles (UAVs). The UAVs extract feature vectors from the received telemetry and combine the feature vectors for cluster analysis of daily trajectory patterns, enabling timely response to the change of vehicle state. We further develop algorithms to fuse user preference information for trajectory design and improve the telemetry collection efficiency, balancing the desired privacy elements so that the cluster discovery quality is maintained despite having fewer telemetry sources. Experiments based on real-world data confirm the effectiveness of the proposed solution.

As for the offloading computation, modern autonomous vehicles have high computation power and significant amounts of memory and storage. The sensing, computing, and communication capabilities of these computing platforms are already advanced and will keep developing. These conditions make our proposed method applicable and practical to a broad and ever-growing range of scenarios for privacy-preserving.

In this section, we present the design of a privacy-preserving framework for data collection of autonomous vehicle telemetry. We propose a mobile edge computing assisted LCP-based framework to address the privacy issues. We make use of newly emerging mobile edge computing technology. Unmanned flying vehicles are deployed or equipped with MEC servers to form a flying ladder. The MEC server is responsible for uplinking the offloading computations of nearby autonomous vehicles and then uploading all the data to the data center when flying back. This framework has some appealing properties, such as low transmission cost, and can ensure the privacy of the data uploading process.

6.1. Architecture Overview

We provide reasoning about the value of k and α parameters for data anonymization. In the case of automobile telemetry, the number of affected entities and the cost of k may vary and should be evaluated per message. We also explain and compare per-message traditional and differential privacy probability distribution fitting methods.

Then, we implement a software tool that applies these privacy protection solutions. In order to assist the user, the tool provides a prediction of the total number of affected entities considering each possible privacy threat. It shows contextual information about the road type mentioned in each message, and it estimates the cost of using different k-anonymity building block measures to increase the amount of noise introduced into the manipulated telemetry.

We consider a public data ingestion authority of a transparent data vendor as sufficiently different stakeholders, and we develop privacy threat models for these stakeholder organizations. We define the privacy protection solutions for each privacy threat and propose privacy-preserving telemetry message generation frameworks.

Second, we propose and implement a software tool that operates based on quality-specific indicators and is capable of cleansing telemetry messages. This tool is being released to the public to lay the foundation for future development of privacy-preserving surveillance software.

First, we provide insights into the stakeholder ecosystem and provide novel observations and analysis of vehicle driver passengers. We perform a thorough risk analysis and offer a balanced choice of privacy threats.

We propose privacy-preserving data collection frameworks for autonomous vehicle telemetry. They enable organizations processing telemetry to avoid the regulatory risks of dealing with personal data by providing them with access to richer, yet semantically valid, datasets that are not in their original form.

6.2. Key Components and Functions

The system presents identity-protection mechanisms that do not require any identity, attribute, or functionality registration. Components are both low-cost and flexible and there is no need for any preliminary input to the system when sharing any attribute with multiple potential data users. The use of such a system allows peers to achieve the coalition ascribed system properties whereby a data attestation authority can accomplish denial of service, integrity, confidentiality, and service-level coalitions without common control or virtually any other prior collaborative interoperability. Anti-traffic analysis protections and distributed system recovery of functions apply to high-frequency or high-velocity intermittent-link communications such as keyless engines, communications among vehicles or drivers, hard-braking notification, stop sign, or red-light notifications, hazard avoidance information, or headlight sighting.

6.2.1. In-Vehicle

The proposed framework provides privacy for all entities that are involved in ADAS or AD vehicle operation. We define the roles of the components and entities that orchestrate the data sharing and networking within the context of a vehicle or across the fleet. Our goal is to give a fine-grained understanding to support the careful design and engineering of tools and techniques into operational and production systems. It is evident that to deployed real systems, such as train sub-systems, make extensive use of the technologies and concepts that are detailed in this and other privacy-centric publications. In the discussions and descriptions that follow, we refer to the entity and performance attributes within the context of our framework. Safeguarding information flow between entities, isolating sharing patterns, handling attorney-client privilege, and being under user or driver controls and consent are highlighted with special emphasis.

7. Implementation and Case Studies

We implemented PASEA on Robot Operating System (ROS) (ROSRP 2020), a typical middleware for robotics. Two separate nodes that are connected by the same wifi access point form a ROS network. A ROS node can use PASEA's APIs to shield the required sensitive telemetry. PASEA delegates the activity of the infrastructure server to the PPLHTM, which is widely used in the RSUs and the edge server in the case studies. PASEA provides a micro-service consisting of the parameters shown in Table 2, which indicates that the user's request consists of the RoC, CoT, and a number of parameters for filtering. The response of PASEA includes the data that satisfy the request, the RTT, and the disabling flag that allows us to verify PASEA's capability by monitoring the CoT.

In this section, we report the implementation of PASEA, the environment setup of the integrated data collection frameworks, several sets of transmission efficiency-related experiments, and finally case studies that leverage the data collected via our TCFs to detect, diagnose, supervise, and forecast abnormalities of the sensory or transmission parts of the AV operational telemetries. To evaluate the localization performance when we replace the infrastructure server with an aggregated server, the joint accuracy is also considered.

7.1. Simulation Environment Setup

We developed our simulations on the basis of the OpenDS (Open Data Store) and OpenCAV (Connected Autonomous Vehicles) simulation environments. The OpenDS data layer provides storage and processing of metadata, raw data, and multidimensional data. Both OpenDS and OpenCAV run traffic and place related queries and interpret the output in a visualization tool. Then we modified these two simulation environments to include our telemetry version and form the simulation engine used in this section to evaluate the privacy-preserving query processing algorithms and the machine learning predictive model application in the context of autonomous vehicle telemetry.

We create a simulation engine for our privacy-preserving query processing algorithms and the machine learning predictive model application in autonomous vehicle telemetry in simulation. In this simulation engine, we consider sequencing data collected in vehicular networks, building queries that ask for results on the subsequent data items obtained from the vehicle-to-cloud communication and vehicle-to-vehicle communication, querying location-dependent and aggregated query results from the map store, querying classification models from the model store to process access and violation privacy queries, and evaluating the predictive model application results.

7.2. Case Study 1: Urban Traffic Monitoring

Nevertheless, the high-quality data produced by traffic monitoring systems cannot be shared due to privacy concerns. For example, although images or videos might be displayed in the public information services of urban areas, these images or videos cannot be used to collect and analyze sensitive information, such as the private appearances of urban people, cars, or other social objects. To overcome the privacy-preserving challenge of urban traffic monitoring, current researchers first anonymize the sensitive data via state-of-the-art privacy-preserving algorithms and then output the processed data in the public service. Therefore, traffic monitoring data sharing becomes a typical privacy-preserving data sharing problem.

Traffic monitoring has become a critical capability of smart cities. This capability can help improve other urban systems, including emission control, energy consumption, resource allocation, public safety, and traffic management systems. With the maturity of machine learning and computer vision technologies, traffic monitoring of urban areas has made a significant leap. Nowadays, traffic monitoring systems have the capability to output real-time urban traffic statuses, such as the spatial-temporal distribution of the flow, speed, trajectory, and so on. Several types of data are collected by traffic monitoring systems, including images, videos, sounds, and LIDAR point clouds.

7.2.1. Introduction

8. Evaluation and Performance Metrics

We evaluate our system using a real-world dataset collected from the Tesla Model S vehicle. We first examine the performance gain due to the employed query optimization algorithm and the benefits of the proposed thick-client-client-optimization in reducing the communication cost between the thick clients and the user-level clients. We then compare the performance of the proposed CHEMA and the state-of-the-art dual-cypher RRPHE encryption scheme (ES) for the scenario of vanilla (joined) reports production. We quantitatively demonstrate that CHEMA outperforms ES by an order of magnitude. As the thick clients are employed by the Reporting Server, the reported overhead for these shielded user-level clients in terms of memory, CPU cores, and storage is only 3%, less than 1%, and 50% of those resources respectively. We also justify the observed tiny latencies introduced by the thick clients and we see that most of the communication is of SQLite meta-statements nature. We choose the April 11, 2018 dataset released by Geiger et al. collected by a Tesla Model S and the Amon research platform. This dataset contains a total of 9,500 CSV files, and each file corresponds to the telemetry transmitted by the vehicle every second. As Geiger et al. promised to delete the dataset over time, we respectfully choose not to share the dataset. In order to simulate the full report generation, we slightly adapt the requested attributes in the following scenarios and run the reporting server jobs using the Tuleap 7.11.99 mapping rule and split standard reports packets into three user-level clients.

8.1. Privacy Metrics

In the context of personal privacy, organizations and data custodians routinely apply statistical tests or anonymization models to create privacy meshes of sufficient density to prevent data infrastructure attacks or induced re-identification attacks that could hypothetically lead to a Deanonymization (or linkage) attack. Such privacy processes are to a large extent now commonly taught and are largely well understood. More difficult, however, is the application of these "rules, limits, and expectations" to aggregated or derived data, as we discuss later in this chapter. Such data aggregation poses a significant data privacy challenge, particularly in the case of Highly Automated Vehicles (HAVs) where large amounts of data are created by modern sensors and systems. These commercial or company-held autonomous vehicle (AV) datasets grow large over time, leaving a significant data footprint. Many such datasets are not anchored to real-world parameters; rather, they employ derived agglomerated data aggregated from large bloated clusters of inferred ride snapshots, from which it is often very difficult for some data subjects to opt out. In this consolidated work, we present privacy-preserving data collection frameworks based on multi-alternatives to data aggregation from real-world datasets obtained over a large fleet of miles of HAV travel and illustrate how real-world metrics can be used to proxy driver and data subject intent for site selection of where the limits of this data collection should lie.

While the concept of privacy is intuitively known by individuals when their privacy is breached, a definition of what constitutes privacy is, however, not so easy to stipulate. An individual's desire and intent for privacy is often based on cultural norms, but at a more conceptual level, the issue of privacy is one of control. This control can be depicted as an individual's right to select what personal information they want to reveal and what they want to keep concealed, a concept commonly referred to as "data privacy." From an operational perspective, privacy can be approximately quantified based on the degree of identifiability in a piece of private information, such as a natural person's inalienable right to keep their private data undisclosed. There are well-recognized legal frameworks that address this fundamental truth, such as the Common Rule of the federal Policy for the Protection of Human Subjects, the De-Identification standard under the Health Insurance Portability and Accountability Act's Privacy Rule (HIPAA), and Article 4(5).

8.2. Performance Benchmarks

The training latency jointly determines the likelihood, quality, and accuracy of the federated learning model, and has implications on logging, notification, and data collection design. For example, partitioning training schedule or reducing the number of contributors or training rounds. In the open benchmarks, we expose the latency from the update request time of each number of contributors (i.e., 10, 100, 1k, 10k) to their corresponding round number.

The indicators are selected as follows:

In this section, we provide performance benchmarks for three selected real-time MPCS privacy-preserving data collection frameworks: secure two-party computation, secure multiparty computation, and federated learning. These frameworks operate with a time budget of 0.8s. The real-time setting captures the real-world use cases of automotive telemetry, such as near-fault military operations, and enables the ECU to possibly raise alarms and reconfigure data collection strategies when potential privacy leaks cannot be addressed in time. The 0.8s benchmark is currently offered to real-time applications with the need to meet fast reaction times, such as high-speed train operation monitoring and self-driving cars. A time budget reduction to 0.1s would be ideal for upgrading privacy control to real-time mandates, but would require tackling technical challenges such as noise reduction or modeling advancements.

9. Future Directions and Research Opportunities

Reliability of Privacy-Preserving Frameworks under Real-world Complications: The usability and effectiveness of state-of-the-art privacy-preserving data collection frameworks in AVs in the real world depend on the extent to which the frameworks can absorb and tolerate real-world complications such as noisy sensor reading, non-ideal hardware, network instability, and human-introduced exceptions. It can be beneficial to develop new privacy-preserving methods or modify existing methods to address these various complications while still guaranteeing their privacy-preserving properties. In order to maximize the effectiveness of the privacy-enhancement mechanisms, it is necessary to investigate the performance, interpretability, and persistence across the introduced exceptions if additional privacy-preserving techniques are proposed and adopted.

Scalable Privacy-Preserving Data Providers: In real-world cloud-assisted AVs transmitting large amounts of sensory data, it is desirable to have service providers being able to perform privacy-preserving data operations such as aggregate, index, and query over the collected sensory data, without ever decrypting the data. Developing scalable privacy-preserving data provider designs may enable real-time traffic control, scenery change detection, and driving safety evaluation among AV fleets.

Inferring Semantic Information with Encrypted Sensory Data: It is interesting and important to investigate how we can infer contextual semantic information, such as vehicle types and road conditions, from the sensory data encrypted by privacy-preserving techniques, without privacy leakage. These types of inferences are essential for future prevalent data-intensive applications in smart urban ecosystems, but need to be developed carefully.

Metadata Reconstruction: By reconstructing metadata such as control commands, sensor metadata, and timestamps for the events from the utilized data, in addition to the benefits in the downstream driving safety and riding experience applications, we can further direct the privacy-preserving data generation process by using the metadata as side information.

There are several promising opportunities for future work within the wide range of design of privacy-preserving data collection methods and the development of AVs' communications infrastructure. In the following, a few research directions are grouped based on the level of abstraction and generality.

9.1. Emerging Technologies

Collision-detection and pre-brake actuation systems process data from six or more different sensors. If the telemetry data generated by processing these sensors were to be research-collected, the data can be privacy-protected using the emerging technologies in the field of homomorphic cryptographic applications which can not only avoid data privacy breaches during the collection process, but also enable data privacy for after-data market research use. It is very complicated and potentially invasive to mask the personal identifier(s) from all different sensory data and then remove segmentation artifacts before the concatenated data can be privacy-insensitive. The complexity of each task depends on the various dimension of the data. Nevertheless, there are currently available tools to protect personal privacy while collecting communications metadata.

We have several upcoming and some nascent technologies that can effectively provide vehicle-generated sensory data privacy. Sensory data generated, processed, and archived by modern automobiles include inputs to advanced driver-assistance systems such as lane-keeping support, adaptive cruise control systems speed and steering inputs, and pre-crash sensing inputs. Data such as steering wheel adjustments, brake inputs, external airbag classifications, front camera video inputs, and other sensory inputs may also be processed and archived. Modern road operation vehicles such as autonomous trucks and shuttle buses also require similar privacy solutions.

9.2. Interdisciplinary Collaborations

Qualifying control is indispensable for the use of the system variables measured or estimated online which are input to the control design, for example, the voltage values, the magnitude of the currents, and the output. Although the topic of data-driven control is of great interest, the use of system variables in real-time allows the control action to be computed.

In the present case, the MATLAB deep learning toolbox was used for the creation of the neural network described in Section 7 because the algorithm for the construction of the topology depends on functions of the Symbolic Math Toolbox. The transfer function approach proposed in Section 8 for the IEEE-39 network is done using the control toolbox in general (control), and this power flow is computed by solving a simple linear algebra model, generated with functions of the parallel distributed computing.

As shown in the previous sections, the proposed framework depends on an interdisciplinary collaboration between the fields of control, cyber-physical systems, privacy, and security. The objective is to preserve the privacy of the collected data while also ensuring that the data is useful for building control algorithms.

10. Conclusion

The main challenge is that the road statistics based on real-time telemetry data are very sensitive, revealing very accurately whether a vehicle is driving on the road. Furthermore, an inaccurate collection could mask a significant number of vehicles from some pieces of road, leading to high inaccuracy of the statistics. The proposed framework aims to solve the above problems. The key idea is to model the trajectories of vehicles, the points-of-interest on the road, and the observed road statistics as a three-layer Markov network, and perform approximate inference with the belief propagation algorithm. The proposed framework has been validated on traffic volume counting and road surface condition statistics, demonstrating its effectiveness.

We present a privacy-preserving data collection framework for autonomous vehicle telemetry. We consider a server that wants to collect road statistics, such as traffic volume, road surface conditions, and traffic signs and signals. To populate its statistics, the server provides a fleet of vehicles with road maps and GPS trajectories over time. Each vehicle computes the road statistics along its trajectory, adding some random noise, and reports the results to the server. At the same time, the server adds random noise to the road statistics from the input GPS coordinates and thus helps to obfuscate the source of the observed trajectory. The server's goal is to perform accurate road statistics while protecting the vehicle's privacy.

10.1. Summary of Findings

Autonomous vehicles continuously capture multivariate sensory inputs for real-time decision-making. The significant amount of highly contextual real-world conditions (adverse telematics) needs more adversaries' responsibility assurance than acceptable risk and crime-free use after necessary modifications (motivation underpinning industry standards and mandates). Independent trusted third parties collect various statutory and ancillary adverse telematics for automotive stakeholders in multimodal fields such as post-income tax submission, theft reporting, and fraud countering. Considering the data harvesting privacy implications, the digital exhaust from various sensors may not flow to a trusted third party at the state until further provisions are established for adversary privacy. The research addresses this unmet need and presents a data collection voice in the Privacy by Design (PbD) discussion.

Automotive stakeholders support their claims with vehicle data to meet verification, validation, and trustworthiness requirements. Privacy-preserving data collection frameworks establish stringent privacy and security standards by design. While developing the Adverse Vehicle Telemetry Collection Framework (AV-TCF) for autonomous vehicles and using multiple representative methods across key areas of privacy for chronological data collection, this investigation establishes that various architectures can provide robust protection for each method. The differentiation allows automotive stakeholders to choose a suitable implementation based on their data collection requirements and local constraints, focusing on mandated legislation. The findings reinforce the argument that enforcement is more critical to secure the privacy baseline established by the privacy and security standards.

10.2. Implications and Recommendations for Industry

Ensuring ADS safety integrity and assurance, especially given the size of the driving spaces, needs to result in strategies and solutions that can match the magnitude of impending driving domain instances that will be necessary for end-to-end validation and support the assertions of safety. The corporate decision to not data share seems like an understandable protection measure given potential exposures which may be created by litigation. However, as we move to a world where Universal Design Qualification on which DNN use is predicated, substantive data collection partnerships are both the logical risk management approach, and offers the means to support, in an evidence-driven way, the safety assertions being made. Collecting real world operational DNN-based system data demonstrates and helps provide evidence that a highly advanced system is, in fact, designed and built to be safe with the abundant data it requires. Data facilitates the concept of comprehensive testing; highlighting corner case and rare usage scenarios that away driving, ensuring, that the tested completeness aligns with the design & hardware capabilities, and the residual risk is consistent with safety expectations. Closing this gap is the critical enabler to instilling industrywide confidence in ADS safety performance.

Sharing operational data - that can often contain both vehicle and infrastructure design details - has been identified as a key challenge before the potential benefits of sharing of CV operational data can be realized. At present, understanding that there are adverse implications of sharing operational data (challenges of IP exposure, safety liability, etc.), OEMs and fleet operators will not publicly share the abundance of data likely available from their growing/operating vehicle fleets. Contributing to this dilemma, appreciation of the lack of substantial data and validation evidence from the real world that establish credibility/assurance of DNN-based system safety is concerning automobile industry stakeholders. Using DNNs for perception in perception and acting systems of ADS creates an environment in which distrust of the decision making processes may be encountered by the acceptable safety integrity of these systems.

Reference:

1. Perumalsamy, Jegatheeswari, Bhargav Kumar Konidena, and Bhavani Krothapalli. "AI-Driven Risk Modeling in Life Insurance: Advanced Techniques for Mortality and Longevity Prediction." *Journal of Artificial Intelligence Research and Applications* 3.2 (2023): 392-422.

2. Karamthulla, Musarath Jahan, et al. "From Theory to Practice: Implementing AI Technologies in Project Management." *International Journal for Multidisciplinary Research* 6.2 (2024): 1-11.

3. Jeyaraman, J., Krishnamoorthy, G., Konidena, B. K., & Sistla, S. M. K. (2024). Machine Learning for Demand Forecasting in Manufacturing. *International Journal for Multidisciplinary Research*, *6*(1), 1-115.

4. Karamthulla, Musarath Jahan, et al. "Navigating the Future: AI-Driven Project Management in the Digital Era." *International Journal for Multidisciplinary Research* 6.2 (2024): 1-11.

5. Karamthulla, M. J., Prakash, S., Tadimarri, A., & Tomar, M. (2024). Efficiency Unleashed: Harnessing AI for Agile Project Management. *International Journal For Multidisciplinary Research*, *6*(2), 1-13.

6. Jeyaraman, Jawaharbabu, Jesu Narkarunai Arasu Malaiyappan, and Sai Mani Krishna Sistla. "Advancements in Reinforcement Learning Algorithms for Autonomous Systems." *International Journal of Innovative Science and Research Technology (IJISRT)* 9.3 (2024): 1941-1946.

7. Jangoan, Suhas, Gowrisankar Krishnamoorthy, and Jesu Narkarunai Arasu Malaiyappan. "Predictive Maintenance using Machine Learning in Industrial IoT." *International Journal of Innovative Science and Research Technology (IJISRT)* 9.3 (2024): 1909-1915.

8. Jangoan, Suhas, et al. "Demystifying Explainable AI: Understanding, Transparency, and Trust." *International Journal For Multidisciplinary Research* 6.2 (2024): 1-13.

9. Krishnamoorthy, Gowrisankar, et al. "Enhancing Worker Safety in Manufacturing with IoT and ML." *International Journal For Multidisciplinary Research* 6.1 (2024): 1-11.

10. Perumalsamy, Jegatheeswari, Muthukrishnan Muthusubramanian, and Lavanya Shanmugam. "Machine Learning Applications in Actuarial Product Development: Enhancing Pricing and Risk Assessment." *Journal of Science & Technology* 4.4 (2023): 34-65.