Making private GPS data available to policy makers: Investigating the feasibility of multi-party computation for smart mobility

Marina Wiemers , Zekeriya Erkin
Delft University of Technology
Department of Intelligent Systems, Cyber Security Group
Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands

Abstract

With recent advances in performance and complexity, multi-party computation, a privacy-preserving technology which allows for joint processing of hidden input data, has lately been found to be applicable in a number of use cases. Despite existing implementations for secure data aggregation, substantial adoptions of the technology remain limited in the industry, in particular within the domain of smart mobility. This paper addresses the current issue of the mobility data shortage by investigating the potential and feasibility of multi-party computation to share data with policy makers, and proposes a solution based on additive secret sharing. On the basis of a literature study and interviews with infrastructure management authorities, as well as micro-mobility service providers, the drivers of, and barriers to employing a secure data aggregation scheme were identified. The results suggest that the technical solution appears feasible given existing implementations, while trust, acceptance and willingness of participants emerged as obstacles to a realisation.

1 Introduction

With the initiation of micro-mobility providers, i.e. companies offering services such as e-scooters and bike-sharing, the potential for smart mobility within cities has seen a sudden increase in the past decade [1]. This shift in transit from privately owned vehicles to shared means of transportation has not only sparked discussions in terms of sustainability and accessibility [2] but has also introduced a surge of personal, geolocational data [3]. Simultaneously, the practices of data-driven decision making are being explored by public authorities, allowing for a quantitative analysis and validation of policies via concrete data [4]. Within the domain of mobility, applications range from giving optimal travel advice, to crowd control, to infrastructure decisions [5], ultimately aiming at inciting sustainable economic growth and greater life quality [6].

While open mobility data projects both in and outside Europe do exist, lacking quality, standardisation and governance of the data are only some examples of current limitations that public authorities are facing [7]. In response to this, initiatives such as the City Data Specification - Mobility (CDS-M), a data standard defined by a group of Dutch cities in March 2021, are aiming to facilitate the exchange of data between authorities and mobility providers [8].

Nevertheless, there still remain issues beyond establishing data standards, namely the privacy and protection of mobility data as a study commissioned by the Municipality of Amsterdam has identified [9]. With the introduction of the General Data Protection Regulation (GDPR) in 2018 [10], the extent of sharing identifiable data, such as user profiles, geolocational data, and vehicle information with third parties has been limited due to factors such as transparency, availability and trust [11] [12]. Moreover, cases such as the re-identification of seemingly anonymised taxi cab data in New York City [13] and further investigations on the uniqueness of mobility traces [14] have raised concerns regarding how identifiable trip data can be. Several studies have already explored the impact of the GDPR on aspects of smart cities, demonstrating the importance of techniques such as anonymisation and Privacy by Design to employ strategies that take the security and safety of user data into account at the design level [15] [16]. Furthermore, the micro-mobility industry and its data collection can be described as both, cooperative and competitive; although data holders can extract useful insights from combined data, they are still reluctant to provide their own to competitors. Thus, service providers also express a limited willingness to share explicit data with public authorities, despite potential benefits of collaboration, due to commercial sensitivity and limited trust [17].

One potential solution for private data collaboration beyond license agreements is multi-party computation (MPC). This privacy-enhancing technology allows multiple parties to process secret input data via cryptographic protocols without disclosing the individual inputs [18]. A simple and illustrative example thereof would be determining the individual with the highest salary within a group while none of the participants ever disclose their income to each other. Via encrypted communication between the participants and random partitions of the their data, operations such as aggregation, statistical models or even voting schemes can be employed in a secure, privacy-preserving manner. While MPC has not been explored in the field of trip data analysis, other instances in related domains have been proven effective al-

ready, demonstrated in cases of smart grid optimisation by monitoring utility consumption [19] or establishing improved congestion pricing [17].

The administration and management overhead of MPC, however, are major obstacles of the technology's employment due to connection instability, trust establishment, and the heterogeneity of data [20]. Furthermore, concrete and practical use cases of MPC are only starting to be explored, leaving the majority of the industry still unaware of the technology [21].

The aim of this research is to investigate the potential of MPC as a secure solution for mobility data aggregation. By means of the use case of sharing geolocational trip data for occupancy reports, we propose a theoretical architecture to allow micro-mobility providers to contribute their user's GPS data to large-scale data analysis. With the use of MPC techniques, we present a way of giving decision makers access to company-owned aggregated data from a wide range of sources to evaluate infrastructure and traffic management improvements.

The remainder of the paper is structured in the following way: Section 2 illustrates the use case in more detail and provides building blocks necessary to understand the applied MPC techniques. The methodology of this research is presented in Section 3, conveying the use of a literature review and stakeholder interviews to elicit requirements which are discussed in Section 4. The design of the devised architecture and elicited requirements are presented in Section 5, followed by an in-depth discussion of advantages and limitations in Section 6. Sections 7 and 8 reflect on the performed research in terms of ethics, reproducibility and main insights, concluding with identifying potential future work.

2 Background

The matter of aggregating data in a secure manner with the use of multi-party computation has already been the subject of various research papers. This section provides an overview of related works and establishes the preliminaries for MPC, additive secret sharing, and possibilities of statistical analysis on aggregated mobility data.

2.1 Multi-Party Computation

First formulated in the 1980s by Yao via the *Millionaire's problem* [22], multi-party computation is a cryptography field which takes distrust and adversarial, or corrupt, behaviour among the participants into account. It describes a system in which a set of n parties $P_1, ..., P_n$ compute some function $y = f(x_1, ..., x_n)$ on a combination of each parties' private input x_i [23].

Generally, the complex security requirements of an MPC protocol are defined by the *Real/Ideal Simulation Paradigm*, allowing a reduction of the system to more easily understood system [18]: An ideal world is envisioned, where a trusted uncorrupted party receives all the participants' inputs through secure communication channels, evaluates a given function, and reveals the output of it. By comparing the *real world* implementation with the *ideal world* scenario, the paradigm provides a basis with easily verifiable conditions for deeming whether an MPC protocol can be deemed secure. Thus, emerging security properties include:

Input privacy: The individual inputs should remain private and only information derivable from the output should be accessible.

Correctness: The output should be guaranteed to be correct.

Independence of inputs: The input of corrupted parties should be independent to the honest parties' inputs.

Guaranteed output delivery and Fairness: Access to the output should be guaranteed to all parties and not be able to be tempered with.

Furthermore, a substantial assumption of MPC protocols are differing models describing the extent of adversarial powers. *Semi-honest* or *passive* adversaries are assumed to follow the protocol as defined. However, they try to gather information about the remaining parties by keeping a log of all sent messages. *Malicious (active)* and *covert* adversaries, on the other hand, may behave differently to the protocol specifications to attack or break the system, calling for more strict security requirements.

2.2 Privacy-Preserving Data Aggregation

As numerous prior work has shown, MPC lends itself as a feasible and viable solution to aggregate data from several sources without compromising input secrecy. Notable cases concern themselves with linking medical records from different healthcare institutes, for instance the BigMedilytics pilot study aiming at identifying heart failure patients [24], or the analysis of genomic data via crowdsourcing, as proposed by Cho et al [25].

Similarly, a large-scale statistical data analysis was conducted in 2015 using the Sharemind framework. In collaboration with the Estonian Center of Applied Research, the correlation between students with a job and those finishing within nominal time was investigated using MPC [26].

A further well-researched application of MPC are smart grids where various techniques have been proposed to monitor consumption and detect potential leaks [27]. Kursawe et al. [19] elaborated upon this idea, introducing four different protocols. Each of these offer aggregation and comparison to pre-established norm values by adding random masking values to conceal the real meter readings.

One commonality between many of the aforementioned protocols is the division of participating parties into their different functionalities. While some cases, such as [19], suggest a peer-to-peer network offering a fully decentralised system where each party fulfills all functions, other instances propose a client-server model, allowing for reduced communication overhead and costs [26] [28]. These roles are defined as follows:

Input parties (IP): Participants which contribute the input data for computations

Computation parties (CP): Parties which evaluate the function, such as a third trusted party or servers

Result parties (RP): Stakeholders which obtain the computation's result and use it for further applications

2.3 Additive Secret Sharing

The secure computation technique that is commonly used for privacy-preserving aggregation or clustering of horizontally partitioned databases is known as *secret sharing* where a secret input value s is divided into a predetermined number n of secret shares s_n. The scheme consists of two mechanisms, namely a function

$$GenerateShare(s) \rightarrow (s_1, ..., s_n) \tag{1}$$

which creates n random shares, and a recovery function

$$Recover(s_{i_1}, s_{i_2}, ...s_{i_k}) \rightarrow s \tag{2}$$

that uses at least k shares to reconstruct the secret [23].

For instance, in Shamir's secret sharing the shares are generated by producing a random polynomial q, of degree $k-1$ where every share holder i evaluates their share using $s_i = q(i)$ [29]. To reconstruct and obtain the secret, following Eq. (2), at least k participants send their shares to one another and interpolate them use Lagrangian interpolation. In this *k-out-of-n sharing scheme*, $k-1$ shares have to be independent in order to reveal no information of s.

One particular instance of secret sharing is known as *additive secret sharing*. In this scheme the individual shares sum up to the secret, relying on *additive homomorphism*, a property defined such that for a function h and inputs u and v

$$h(u+v) = h(u) + h(v) \tag{3}$$

.

Thus, the sum c of two secrets, a and b, can be performed in a distributed manner, such that

$$c = \sum_{i=1}^{i=n} c_i = \sum_{i=1}^{i=n} (a_i + b_i) = \sum_{i=1}^{i=n} a_i + \sum_{i=1}^{i=n} b_i = a + b \tag{4}$$

.

Each party i adds their respective two shares, a_i and b_i, and sends this sum to the other parties. Subsequently, each party sums the received added shares to obtain the result.

2.4 Statistical Analysis on Aggregated Mobility Data

Given the issue of potential re-identification of individuals in mobility data [14], different approaches to mitigate the risk have been discussed in recent years. While Feng et al. have suggested the use of federated learning to predict human mobility while protecting user privacy [30], other methods include large-scale models which only convey cumulative information, such as heat maps [31] and clustering methods [32].

Given the scope of this research and to serve as a proof of concept, the suggested analysis on mobility data in this paper is limited to elementary queries, such as linking and aggregating trip data based on a specific location and timeframe. However, these fundamental queries open the door for more sophisticated statistical models, such as travel behaviour and route choice models. Further extensions of the proposed system are elaborated upon in Section 6.

3 Methodology

To arrive at a sensible and feasible design for a privacy-preserving data sharing architecture, the process of this research was split into two main phases. These are a literature review and interviews with experts and stakeholders within the Dutch mobility industry. The insights thereof were utilised to identify emerging knowledge gaps and potential applications.

3.1 Literature Review

The research was initiated with a study of existing works of multi-party computation. The aim was to gain an understanding of the various implementations to serve as a basis for devising an architecture to the specific use case. This was done by reading up on published journal articles and conference papers illustrating technical and practical operations of MPC, as well as proofs conveying the integrity of different implementations.

Furthermore, the literature review served as an investigation of existing real-world implementations of MPC, their applications, and an analysis of their feasibility. By consulting the primary sources, different protocols and their technical requirements were compared.

Moreover, the literature review provided insights on current practices within the domain of data-driven policy making. Different use cases and existing models of utilising mobility data for infrastructure investments were examined via reports on pilot projects and research studies focusing on open mobility data.

3.2 Expert and Industry Interviews

To establish the status quo in the domain of smart mobility and collection of data within, interviews with a range of experts and stakeholders, as specified in Table 1, were conducted. These interviews were held in a semi-structured way; organised on the basis of guiding questions, yet leaving room for the respondents to contribute own additions and concerns. Furthermore, the responses were recorded for transcription purposes, and analysed via the open coding method, a qualitative analysis to extract specific feasibility requirements against which the devised system could be assessed.

Experts from independent research institutes and consultancy firms, focusing on the realisation of smart mobility in The Netherlands, were consulted to formulate requirements of a potential data sharing solution, based on their expertise, previous research and experiences with pilot projects. By talking to public authorities and EU-commissioned mobility initiatives, the availability, drivers of, and barriers to shared mobility data at the status quo were obtained. Lastly, technical conversations with micro-mobility providers concerning their data collection and potential collaboration with policy makers were used to validate and evaluate the proposed architecture.

Due to the limited number of interviewees, only a qualitative analysis of the findings was made. The question guide to these interviews can be found in Appendix A.

Organisation	Description
Ministry of Infrastructure and Water Management	Dutch government
Innopay	Consultancy firm
POLIS Network	European Urban Mobility Network
Argaleo	Data-Driven Digital Twins
Mobidot	Mobility ICT service provider
Bolt	Micro-mobility provider
Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek (TNO)	Independent research organisation

Table 1: List of interviewed stakeholders

4 Requirements for Data Sharing in the Mobility Industry

By means of conducting the interviews with stakeholders in the micro-mobility industry, we identified a range of requirements for a potential design. Extending functional and already defined conditions of MPC protocols, such as security, input privacy and robustness, the following describe a viable scheme for collaborating on geolocational data.

Scalability. Given the increasing number of micro-mobility providers in the market, both a scalable technical solution and data sharing business model needs to be put in place. One important aspect in this is the time and computational complexity; in order to run complex analysis on large amounts of data a compatible implementation has to be chosen. Simultaneously, three of the interviewees pointed out the need of a feasible business model which defines the practice, access and cost of using the data aggregates.

Accessibility and Usability. Other evaluation criteria of feasibility are how useful and accessible the insights resulting from the aggregation are. For instance, some use cases rely on the provision of real-time data or whole user journeys which are more difficult to extract from an accumulated source.

Trust. Due to the competitive nature of many mobility service providers, there is little readiness to grant access to collected information to research institutes and public authorities. One factor to mitigate that sentiment is increased trust in the technology itself as well as participating parties.

Privacy and GDPR Compliance. The case of a GDPR-compliant system has already been discussed in previous sections. However, an additional reason declared during the interviews was that some data providers were not willing to share data, using privacy as an "overly protective excuse" [1]. Thus, privacy needs to be ensured by the system.

[1]Anonymous Stakeholder (2021, May 31). Personal communication [Online Interview]

Data Standards. Yet another standing issue is the fragmentation and incoherence of data. Due to inconsistent formats as well as vertical silos, the matter of data fusion has addressed the need for standardised data formats to facilitate operations and ensure integrity.

Acceptance. Lastly, a concern that was raised by all stakeholders was acceptance, both in public and private terms. On the one hand, users of micro-mobility services need to consent to the use of their travel information. On the other hand, beyond agreements and licenses between public authorities and data providers, there have to exist incentives for service providers to participate in the first place.

5 A Privacy-Preserving Architecture for Data Aggregation

This section illustrates the proposed architecture to utilise multi-party computation in the context of aggregating mobility data and offers an overview of how an established framework based on secret sharing, such as Sharemind [23], can be adopted for the purpose.

5.1 Motivating Use Case

For the sake of providing an easily understood example to demonstrate how MPC could be used to privately aggregate travel data from multiple sources, the use case applied in this section concerns an analysis of occupation at arbitrary locations. With the presence of already collected and reliable mobility data in areas of limited traffic sensors, such as pedestrian zones, a quantitative analysis allows for activity monitoring. In order to do so, the aggregate database is queried based on a specified (range of) location and time, resulting in a sum of vehicle instances.

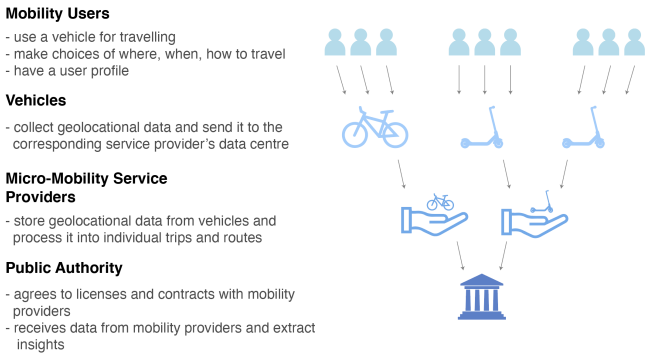

Figure 1: The flow of geolocational data from individuals using micro-mobility vehicles to public auhorities for policy-making

Figure 1 illustrates how policy-makers are currently accumulating information from individuals using open data models; end users of mobility services use vehicles, such as bikes or e-scooters, to travel along a route and time of their choosing. Synchronously, the vehicles are equipped with sensors, collecting among others the longitude, latitude, date and timestamp of the carrier's current location at a specified frequency. This information is then sent to the corresponding service provider which in most cases stores it and often runs

their own analyses, to improve for instance their supply-and-demand model. Finally, this data is shared with public authorities in a few different ways. In some cases, service providers choose to report only high-level insights while others include a third party or an API to make pseudo-anonymised trip data available. Moreover, this step usually includes a range of licences and agreements specifying the exact use of data, as well as payment models.

5.2 Security Assumptions

For this use case the *semi-honest* security model is assumed. As briefly described in Section 2.1, all participants are described as *honest-but-curious*; they keep track of all messages and outcomes in an attempt to learn more information about the remaining parties. By employing this assumption, it is guaranteed that no inadvertent information is disclosed during the computations [18].

The reasoning for this rather lenient model is twofold. Firstly, service providers can potentially profit and gain market power via learning from their competitors' user travel behaviour. Thus, they can be assumed to be honest-but-curious, whether they are merely data providers or also computing parties and should only be able to see encrypted shares. Secondly, the existence of agreements and contracts between services providers and authorities allows us to argue for an honest behaviour among participating parties.

5.3 Architecture Overview

Following the model proposed in [28] and described in Section 2.2, the stakeholders are divided into different entities based on their function, economising performance as only a subset of parties is included in the computation phase itself. The structure of this architecture is based on the one used in [26], illustrated in Figure 2. Accordingly, the micro-mobility service providers are considered data providers which distribute their inputs via secret sharing among the computational parties. These computational parties are composed of servers, reasoned to be best-performing at an amount of three [23], which authenticate participants, store the data, and evaluate the aggregation function on a selected query [33]. To mitigate costs, adversarial risk and monopolisation, it is advised for them to be provided by different stakeholders. Ultimately, the computing parties send the query outcome to the result party, in this case a traffic managing public authority, which may then perform further analysis.

5.4 Pre-processing, Aggregation and Constraints

A currently prevalent issue in the domain of smart mobility is data fusion, as different mobility providers adhere to different formats. Addressing this issue, constraints on the data are inevitable. While one solution would be to sanitise and transform the data as part of the protocol, the practice of a predefined data standard is more effective, given that they are currently in the midst of being employed [8]. The use case of determining spatio-temporal occupation, such as heatmaps, and current data structures suggest a database schema consisting of five keys, namely the vehicle ID, date, time, and discretised longitudinal and latitudinal values of the vehicle's position.

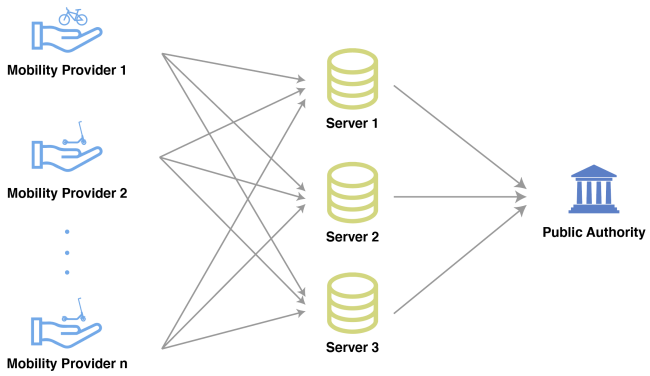

Figure 2: The proposed architecture of participating stakeholders

In order for the three servers, constituting the computational parties, to be populated with the service providers' mobility data, each service provider transforms their data into the agreed upon database schema, hashes the values, and indexes each record, i.e. row, for future reference. Subsequently, each attribute is randomly fragmented into three shares, according to the secret sharing schemes described in Section 2.3. In doing so, a hashed record r with corresponding 5 attributes $r_1, ..., r_5$ results in a partition of $r_{11}, r_{12}, r_{13}, ..., r_{53}$. An attribute r_i is split into its shares r_{i1}, r_{i2}, r_{i3}, such that

$$r_i = r_{i1} + r_{i2} + r_{i3} \tag{5}$$

. Lastly, the generated shares are accordingly distributed among the three servers and stored in there in their hashed and partitioned state, such that each server j has a 5 local attribute shares for each record, $r_{1j}, ..., r_{5j}$. An example and proof of privacy of this method is provided in Algorithm 4 in [34].

5.5 Querying and Data Retrieval

Once the data has been securely distributed, the result party can query a request to define the operation to be performed and its parameters. In a comparable manner to conventional database management systems, an application serving as an interface between the computational and result parties, such as Sharemind's environment Rmind, translates the request and forwards it to each of the three computational parties. Each server then performs the operation, depending on the type of query within one or multiple communication rounds, for instance when comparing values, as outline in Algorithm 1 in [26].

Before the outcome fragments are shared with the result party, an important constraint could be to not disclose the result if the outcome of a query potentially reveals too much, for instance the count of vehicles for a specified time and location is below a certain threshold. Thus, to mitigate the risk of re-identification, a remedy such as a prompt to choose a larger range in the spatial or temporal domain should be implemented.

Finally, the public authority is able to retrieve the result of the query by interpolating the outputs, i.e. adding the shares, provided by the three computation servers.

6 Discussion and Feasibility Analysis

In this section, the proposed architecture is validated by means of technical and non-functional requirements. Elicited via prior works and industry research, these requirements form varied criteria for the feasibility of the scheme.

6.1 Discussion of Technical Feasibility

Privacy and Security. The key aspects when assessing a privacy-preserving technology are the input privacy, security and robustness of the scheme. As defined in Section 2 and elaborated upon in the architecture, privacy of the proposed architecture is accomplished as a result of the employment of additive secret sharing. Working under the assumption and fulfillment of a *honest-but-curious* security model, both the input privacy and robustness of the system remain intact as long as less than the majority, i.e. maximally one computational server, is corrupted. Despite adversarial attacks as such seeming unlikely, considering the presence of license agreements and legal implications, it should be taken into account that the applied model is a rather lenient one and does not account for complete security.

Scalability. The proposed solution's scalability can be evaluated in terms of communication overhead, computational complexity and latency. Due to the use of only three servers, in contrast to letting data providers performing the computations themselves, the necessary soft- and hardware is limited to a small subset, diminishing the need for time-costly communication and key distribution [23]. Solutions similar to the design suggested in this paper, such as Sharemind, show a wide range in performance; as [33] summarises, integer arithmetic operations extended from 2 minutes to more than 5 hours after optimisation, depending on the amount of data.

As these figures suggest, a similar implementation adapted to the case of aggregated trip data is within reasonable bounds. However, such a utilisation lends itself more for the analysis of historical data, for instance studies that investigate the change of mobility patterns over a range of months, rather than applying models to real-time data.

Nevertheless, it should also be noted that the proposed design is only a theoretical solution as it was conceived without estimates of amounts of data and computational complexity of potential queries. As follows, an appropriate analysis of the sharing scheme for this particular use case remains open.

6.2 Discussion of Non-Technical Feasibility

While technical requirements are considerably easily verifiable through prior work in the field, the assessment of non-functional feasibility criteria was performed by consulting experts, resulting in the elicitation of requirements as presented in Section 4.

Data Standards. Similar to most data-driven mechanisms for statistical analysis, our design necessitates a coherent data format to accurately perform operations. While not employed universally, standards for mobility data are being established and introduced to the market, and can thus be seen as both, a driving and limiting factor to the feasibility of the system. Yet, governance posed itself as a major factor contributing to the upkeep of not only data sharing schemes but also new standards for emerging technologies such as MPC.

Accessibility and Usability. Currently, there are only a limited number of instances where data-driven policies are employed for infrastructure decisions. Due to the topic, especially in combination with smart mobility, being rather novel, the type of statistical models, amount, and frequency of data required to perform robust analyses remain undetermined.

Concerning the usability and employability of the system, one advantage is the present collection of data itself as almost every micro-mobility provider stores at least some information on their users' trips. While researchers often face the issue of gathering large amounts of information, the data in this use case already exists, including users' consent to store and process it.

Trust. Despite MPC taking distrust among parties into account, the interviews identified trust outside the system as a major barrier, illustrating examples of data providers being uncooperative when asked for collaboration. Combined with the inevitable skepticism and doubt towards most novel technologies, this distrust might lead to decreased participation and delayed employment of the protocol.

GDPR Compliance and Commercially Sensitive Data. As discussed in Section 6.1, by adhering to input privacy and robustness, the proposed solution ensures that the inputs supplied by the service providers remain secret. Moreover, the additional constraint, to only reveal outputs to queries if a certain threshold of the sum is achieved, is another privacy measure to prevent the identification of individuals based on their trip data. Hence, it can we argued that the MPC protocol achieves GDPR compliance as no personal or identifiable data is ever shared or processed by third parties.

A similar argument can be drawn concerning commercially sensitive data; neither the result parties nor other data providers are able to extract complete data from individual service providers. Nevertheless, the insights extracted from the aggregates can reveal sensitive information that market players might prefer to keep undisclosed.

Acceptance. Tying into the concept of sharing sensitive data, our proposed architecture remains under the question of acceptance. As various stakeholders in traffic and infrastructure management have pointed out, some form of compensation for the data providers to participate in the sharing scheme needs to be in place. For instance, the query issuing authority might introduce subsidies for participation, a data business model could be conceived or other benefits, such as granting access to the aggregated data itself, could be explored.

In addition, a feasible implementation of the suggested scheme calls for a well-designed governance model. Assuming that this way of using mobility data to make informed decisions for traffic and infrastructure management is on-going, rather than a single instance, according laws and administration beyond the technical implementation need to be set in place.

7 Responsible Research

Having discussed the proposed architecture in regards to technical and non-technical feasibility, this section addresses implications of a system which aggregates geolocational user information, and considers the reproducibility of this research.

7.1 Ethical Considerations of Shared Mobility Data

As the applicability of the GDPR suggests, the collection and processing of geolocational data can be subject to privacy infringements. Despite the demonstrated anonymisation and level of security that MPC can provide, there still exists a risk of adversaries forming the majority of participating parties. In such cases of misuse, various risks and ethical concerns for the different stakeholders may arise.

The end users of micro-mobility services are arguably impacted the most, should their data be compromised. Due to working with time-stamped trip data, individuals could be identified via their transit patterns, i.e. daily work commutes. This could potentially pose risks, including infringement of freedom and surveillance, if the data is misused.

Another significant stakeholder affected by an implementation of the proposed system are the mobility service providers themselves. For them, there is a liability for data leakage which could conceivably lead to loss of market power if competitors gain access. Furthermore, the question of ownership of the data - whether it is owned by the end users or vehicle providers - draws attention to considerations of further (commercial) usage and possible exploitation thereof.

Despite the aforementioned benefits of data-driven mobility policy making, the gained access to aggregated trip data may also introduce maltreatment involving public authorities and society. Misuse of the generated insights could induce potential leverage over the mobility industry. For instance, policies could be used to disadvantage specific service providers. In addition, false interpretation of the aggregation result could lead to rendering low-activity neighbourhoods as negligible. In doing so, the potential of aiding policies or investments might go unused. Lastly, another ethical concern is the question regarding whether the data should be allowed to be reused for other purposes, such as in criminal investigations.

7.2 Reflections on Reproducibility

As was pointed out in the methodology description of this paper, a literature study and stakeholder interviews were performed to arrive at suitable evaluation and validation criteria for a feasible aggregation scheme. Both methods convey limitations due to being a small selection of a much larger compendium of resources. The selected research papers were chosen on the basis of published and peer-reviewed work.

Given the small set of interviewees, bias and mis-sampling of the stakeholder representatives was inevitable. It should therefore be noted that several points made in the discussion and analysis of this paper are based on a restricted, and possibly unrepresentative collection of expert opinions. In addition, the mode of semi-structured interviews might affect the reproducibility, given the lack of a consistent set of questions. To account for reproducibility, the questionnaires can be found in Appendix A and anonymised transcripts of the interviews are available upon request.

8 Conclusion and Future Work

With the use of interviews with stakeholders within the smart mobility industry, the status quo and potential of employing MPC to draw insights from joint data sources has been explored. By showing the applicability of established privacy-preserving data aggregation mechanisms to the analysis of trip data, the technical feasibility and acceptable complexity of such a system was demonstrated. Moreover, an architecture and protocol design using homomorphic encryption and secret sharing was proposed to convey the possibility of performing secure computations on private inputs provided by mobility service providers. Privacy requirements and GDPR compliance were proven to be fulfilled with the use of an appropriate security model.

Nevertheless, potential barriers beyond the technical implementation have been identified, an important one being trust. Despite the MPC solution taking distrust among parties into account, the interviews have shown that mobility providers are still averse to contributing data due to distrust towards both, the technology and policy-makers. Furthermore, the industry research indicated that public authorities and traffic management institutions in The Netherlands have only started to explore the domain of data-driven policy-making, implying limited knowledge of which mobility models are deemed useful for analysing geolocational data.

As this paper has suggested a technically feasible solution to aggregate mobility data, there are various potential areas for future work and further development. Firstly, an implementation in combination with a mobility pilot study could be conducted to accurately predict the time and computational complexity of the proposed sharing scheme. Secondly, further research could include an in-depth evaluation of the willingness of mobility providers to participate, providing quantitative evidence of drivers and barriers. Lastly, by demonstrating how existing frameworks can be applied to a specific use case, this paper may encourage the exploration of similar instances where a privacy-preserving data aggregation scheme could be of aid.

References

[1] Grant McKenzie. Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services. *Computers, Environment and Urban Systems*, 79:101418, 2020.

[2] Anna Kramers, Tina Ringenson, Liridona Sopjani, and Peter Arnfalk. AaaS and MaaS for reduced environmental and climate impact of transport. In *ICT4S*, pages 137–152, 2018.

[3] Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira Jr, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26:301–313, 2013.

[4] Daniel C Esty and Reece Rushing. Governing by the numbers: The promise of data-driven policymaking in the information age. *Center for American progress*, 5:21, 2007.

[5] Paweł Gora and Piotr Wasilewski. Adaptive systems for intelligent traffic management in smart cities. In *International Conference on Active Media Technology*, pages 525–536. Springer, 2014.

[6] Andrea Caragliu, Chiara Del Bo, and Peter Nijkamp. Smart cities in europe. *Journal of urban technology*, 18(2):65–82, 2011.

[7] Noor Huijboom and Tijs Van den Broek. Open data: an international comparison of strategies. *European journal of ePractice*, 12(1):4–16, 2011.

[8] Isobel Duxfield. Dutch cities develop new mobility data standard. https://www.polisnetwork.eu/news/dutch-cities-develop-new-mobility-data-standard/, Apr 2021.

[9] Guusje Guusje van der Vossen and Amy Rook. The opportunities, bottlenecks and practical possibilities of a European data standard for shared mobility operators. https://openresearch.amsterdam/en/page/70006/the-opportunities-bottlenecks-and-practical-possibilities-of-a, May 2021.

[10] Council of 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). http://data.europa.eu/eli/reg/2016/679/2016-05-04, 2016.

[11] SHERPA. How data is used in smart cities. https://www.project-sherpa.eu/how-data-are-used-in-smart-cities/, 2020.

[12] David Eckhoff and Isabel Wagner. Privacy in the smart city—applications, technologies, challenges, and solutions. *IEEE Communications Surveys & Tutorials*, 20(1):489–516, 2017.

[13] Marie Douriez, Harish Doraiswamy, Juliana Freire, and Cláudio T Silva. Anonymizing NYC taxi data: Does it matter? In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 140–148. IEEE, 2016.

[14] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1–5, 2013.

[15] Maria Stefanouli and Chris Economou. Data protection in smart cities: Application of the EU GDPR. In *The 4th Conference on Sustainable Urban Mobility*, pages 748–755. Springer, 2018.

[16] Goran Vojkovic. Will the GDPR slow down development of smart cities? In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1295–1297. IEEE, 2018.

[17] Matthew Tsao, Kaidi Yang, Stephen Zoepf, and Marco Pavone. Trust but verify: Cryptographic data privacy for mobility management. 2021.

[18] Yehuda Lindell. Secure multiparty computation (MPC). *IACR Cryptol. ePrint Arch.*, 2020:300, 2020.

[19] Klaus Kursawe, George Danezis, and Markulf Kohlweiss. Privacy-friendly aggregation for the smart-grid. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 175–191. Springer, 2011.

[20] Marcel von Maltitz, Stefan Smarzly, Holger Kinkelin, and Georg Carle. A management framework for secure multiparty computation in dynamic environments. In *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*, pages 1–7. IEEE, 2018.

[21] David W Archer, Dan Bogdanov, Yehuda Lindell, Liina Kamm, Kurt Nielsen, Jakob Illeborg Pagter, Nigel P Smart, and Rebecca N Wright. From keys to databases—real-world applications of secure multi-party computation. *The Computer Journal*, 61(12):1749–1771, 2018.

[22] Andrew C Yao. Protocols for secure computations. In *23rd Annual symposium on foundations of computer science (sfcs 1982)*, pages 160–164. IEEE, 1982.

[23] Dan Bogdanov, Sven Laur, and Jan Willemson. Sharemind: A framework for fast privacy-preserving computations. In *European Symposium on Research in Computer Security*, pages 192–206. Springer, 2008.

[24] W van Haaften, A Sangers, T Engers, et al. Coping with the general data protection regulation; anonymization through multi-party computation technology. In *23rd International Legal Informatics Symposium, 2020.*, 2020.

[25] Hyunghoon Cho, David J Wu, and Bonnie Berger. Secure genome-wide association analysis using multiparty computation. *Nature biotechnology*, 36(6):547–551, 2018.

[26] Dan Bogdanov, Liina Kamm, Baldur Kubo, Reimo Rebane, Ville Sokk, and Riivo Talviste. Students and taxes: a privacy-preserving study using secure computation. *Proceedings on Privacy Enhancing Technologies*, 2016(3):117–135, 2016.

[27] Flavio D Garcia and Bart Jacobs. Privacy-friendly energy-metering via homomorphic encryption. In *International Workshop on Security and Trust Management*, pages 226–238. Springer, 2010.

[28] Assaf Ben-David, Noam Nisan, and Benny Pinkas. FairplayMP: a system for secure multi-party computation. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 257–266, 2008.

[29] Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[30] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. PMF: A privacy-preserving human mobility prediction framework via federated learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–21, 2020.

[31] Mohamed Maouche, Sonia Ben Mokhtar, and Sara Bouchenak. HMC: Robust privacy protection of mobility data against multiple re-identification attacks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–25, 2018.

[32] Yuchuan Du, Fuwen Deng, and Feixiong Liao. A model framework for discovering the spatio-temporal usage patterns of public free-floating bike-sharing system. *Transportation Research Part C: Emerging Technologies*, 103:39–55, 2019.

[33] David W Archer, Dan Bogdanov, Yehuda Lindell, Liina Kamm, Kurt Nielsen, Jakob Illeborg Pagter, Nigel P Smart, and Rebecca N Wright. From keys to databases—real-world applications of secure multi-party computation. *The Computer Journal*, 61(12):1749–1771, 2018.

[34] Dan Bogdanov. *Sharemind: programmable secure computations with practical applications*. PhD thesis, Tartu University, 2013.

A Guide for Semi-Structured Interviews

A.1 Questions for Data Collectors and Providers

Thank your for agreeing to this interview and participating in my research. As mentioned before, I am a third year Computer Science student at TU Delft, in the Netherlands and am currently working on my bachelor thesis. As part of the Cyber Security Group I am investigating the awareness and potential of MPC in the smart mobility industry.

Before starting, I would like to ask for consent to record this call for later transcription. Furthermore, how may I reference the insights you provide in my paper, i.e. by name, organisation, or anonymised?

Introduction

- Could you describe the position of your organisation within the industry of (smart) micro-mobility?
- What is your role at your organisation?

Collection of Mobility Data

- What types of data are you currently collecting and how?
- For what purpose are you collecting the data, for instance user experience, data-driven business models, simulations,...?
- How is this data stored and processed?
- Are there any data or insights you would like to collect but currently aren't able to? Why is that so?
- Are you using third party mobility data or providing your own to third parties? If so, could you elaborate on the types and purposes of such cooperation?
- When collaborating on data, what types of regulations, agreements and standards are you adhering to?
- In case you are currently not providing any data, what are the reasons for doing so, for instance legality, privacy, competition, cost,...?
- Where do you see requirements that would need to be put in place to allow for secure data sharing among mobility providers?

Data Security and GDPR Compliance

- How is the privacy and security of the data you collect ensured?
- How has the introduction of GDPR affected you? Were there any compromises you have had to make in order to adhere?
- Which measures are currently taken to (pseudo-) anonymise collected user data?

Multi-Party Computation

- Are you aware of Multi-Party Computation? *If not, an explanation and example use case of MPC is provided*
- Where do you see potential use cases of MPC in the mobility industry?
- Are you able to identify driving an limiting factors of the technology, for instance in terms of infrastructure, cost, participation, novelty,...?

A.2 Questions for Public Authorities and Experts

Thank your for agreeing to this interview and participating in my research. As mentioned before, I am a third year Computer Science student at TU Delft, in the Netherlands and am currently working on my bachelor thesis. As part of the Cyber Security Group I am investigating the awareness and potential of MPC in the smart mobility industry.

Before starting, I would like to ask for consent to record this call for later transcription. Furthermore, how may I reference the insights you provide in my paper, i.e. by name, organisation, or anonymised?

Introduction

- Could you describe the position of your organisation within the industry of (smart) micro-mobility?
- What is your role at your organisation?

Collection of Mobility Data

- For which purposes are you analysing mobility data? Could you elaborate on a few use cases or pilot studies that you are currently investigating?
- Which kind of data are you currently using and how do you gain access to this data?
- Which types of statistical models or analysis to you apply to the data?
- Are you able to give estimates of the required amount and latency of data for such analyses?
- Are there any data or insights you would like to collect but currently aren't able to? Why is that so?
- When using data from third parties, what types of regulations, agreements and standards are you adhering to?
- Where have you noticed limitations and challenges in gaining access to mobility data, for instance legally, financially, or in terms of availability, privacy, security,...?
- Where do you see requirements that would need to be put in place to incentivise mobility providers to provide data?

Data Security and GDPR Compliance

- How is the privacy and security of the data you work with ensured?
- How has the introduction of GDPR affected you? Were there any compromises you have had to make in order to adhere?
- Which measures are currently taken to (pseudo-) anonymise collected user data?

Multi-Party Computation

- Are you aware of Multi-Party Computation? *If not, an explanation and example use case of MPC is provided*
- Have you investigated other privacy-enhancing technologies to aid the provision of mobility data?
- Where do you see potential use cases of MPC in the mobility industry?
- Are you able to identify driving an limiting factors of the technology, for instance in terms of infrastructure, cost, participation, novelty,...?