Towards Anonymizing Intermodal Mobility Data for Smart Cities

LEONIE ACKERMANN*, University of Bamberg, Germany
MICHAEL MÜHLHAUSER*, University of Bamberg, Germany
ALEXANDRU BURDUSEL, wikimove UG, Germany
MICHAEL FEDERLIN, wikimove UG, Germany
DOMINIK HERRMANN*, University of Bamberg, Germany
STEFFEN HOLLY†, Psoido GmbH, Germany
DANIELA NICKLAS*, University of Bamberg, Germany
DANIEL WOLPERT, wikimove UG, Germany

As cities seek to optimize their resources for a sustainable and livable future, the concept of intermodal mobility has become increasingly important. However, the collection and analysis of intermodal mobility data is complicated by the need for robust anonymization methods, as privacy and security concerns remain paramount. Existing anonymization methods are either mode-specific or so complicated that they deter potential stakeholders. In this paper, we describe a variety of real mobility data sources for our upcoming field study. With that data, we plan to provide insights into infrastructure utilization and transitions between modes of transport. We further identified several anonymization techniques for mobility data to ensure privacy and acceptance among the citizens. To find suitable techniques for intermodal mobility data, we provide insights from our previous experience on anonymization and discuss the practicability of the identified techniques. Our paper highlights the need for explainable anonymization methods tailored to intermodal mobility data that address privacy and security concerns and pave the way for more accessible privacy-compliant solutions.

CCS Concepts: • **Security and privacy → Human and societal aspects of security and privacy**; • **Information systems → Data structures**; • **Human-centered computing → Mobile computing**.

Additional Key Words and Phrases: intermodal Mobility, Anonymization, Privacy, Smart Cities

ACM Reference Format:
Leonie Ackermann, Michael Mühlhauser, Alexandru Burdusel, Michael Federlin, Dominik Herrmann, Steffen Holly, Daniela Nicklas, and Daniel Wolpert. 2023. Towards Anonymizing Intermodal Mobility Data for Smart Cities. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Municipalities face the challenge of organizing mobility in such a way that limited resources are used as effectively as possible for a sustainable and livable future. In this context, intermodal mobility, i. e., the use of multiple modes of transportation within a single journey or trip, is a key factor for providing sustainable mobility options, which is a goal defined in the Smart City Charter [4]. To provide and optimize mobility services, sensitive data such as trajectories and usage data from mobility users is needed [17]. Typically, before such data can be used for analysis, anonymization techniques are applied to prevent re-identification of individual users. To assess the privacy guarantees of the anonymization

Manuscript submitted to ACM

process, researchers came up with several privacy models such as k-anonymity [18], l-diversity [14], t-closeness [12], and differential privacy [6]. However, many anonymization techniques target only particular transportation modes or data types. In addition, the techniques might be so complex that decision makers, i. e., users giving their consent or data protection officers, do not understand them and thus do not allow using the data. Hence, the goal of the project *explanym*[1] is to develop and demonstrate explainable and practical anonymization techniques in the context of intermodal mobility services and to gain insights on conditions under which such techniques are understood and accepted by those affected. For that, we formed an interdisciplinary alliance with researchers from computer science, psychology, industry and end-users from communal mobility, asset tracking/IoT, and hospital management. The main contributions of this paper are as follows: (i) We outline the challenges that need to be addressed to enable large-scale field studies in the context of intermodal mobility services. (ii) We describe the data for our field study on intermodal mobility data. This includes the different data sources, e. g., pedestrian movements, bicycle data, but also data from public transportation and e-scooters. (iii) We describe insights from our previous work on anonymizing mobility data, identify anonymization techniques from the literature, and assess their practicability for anonymizing intermodal mobility data in our field study.

Apart from intermodal mobility data in the Smart City Bamberg, as presented in this paper, *explanym* also investigates privacy-aware tracking of assets in a clinic. The paper is organized as follows: Section 2 addresses the challenges of collecting and anonymizing intermodal mobility data. In Section 3, we present our data sources to collect mobility data during our field study. Section 4 explores anonymization techniques and their application for mobility data, while Section 5 discusses and concludes our work.

2 CHALLENGES

Several critical requirements and challenges arise when collecting and anonymizing intermodal mobility data for comprehensive urban planning and transportation optimization:

Mobility Data encompass a wide spectrum of variables, including different frequencies and granularities, leading to *Diverse Mobility Data Formats*. Moreover, distinct anonymization techniques are required to suit the specific characteristics of each format. For instance, data collected from public transportation differs significantly from data obtained from an E-Scooter provider. Each data source necessitates tailored anonymization approaches, adding complexity to the overall process. To get a holistic picture of urban mobility, intermodal trips must be seamlessly connected. Mobility data comes from a variety of providers, each using their own anonymization methods. While anonymization is critical to maintaining privacy, it can hinder the integration of data from multiple sources. When data is anonymized on a per-provider basis, the important link between intermodal trips can be broken. Therefore, reconciling anonymization practices with the need for coherent, cross-provider *Intermodal Trip Integration* is a major challenge. The effectiveness of many anonymization techniques depends on the availability of a *large volume of data*. In the case of a small city like Bamberg, the challenge of achieving the minimum data volume for certain anonymization techniques becomes apparent. This limitation necessitates innovative synthetic data strategies or alternative approaches to anonymization to ensure the utility and accuracy of the derived insights.

In summary, while intermodal mobility data hold promise for improving urban planning and transportation systems, it is equally important to acknowledge and address the multiple challenges associated with their collection and

[1] https://www.uni-bamberg.de/explanym/ Accessed on: 2023-09-15

anonymization. Successfully addressing these challenges requires a nuanced understanding of data formats, the development of interoperable anonymization strategies, and solutions for working with limited data resources.

3 MOBILITY DATA SOURCES

As a core part of our research, we plan to conduct a three month field study and collect temporal related mobility data. Some data can be obtained by connecting to existing APIs (live data or by retrospective data download) others will be gathered by distributing sensors to volunteers. As we can see from the following description, the mobility data will be highly heterogenous in terms of update frequencies, data quality, or pre-precessing.

Micromobility services. E-Scooter data is collected from micromobility providers[2]) which implement the Mobility Data Standard (MDS) [16] and the General Bikeshare Feed Specification (GBFS) [7] data standards as required by MDS. The MDS standard covers multiple mobility modes, namely micromobility, passenger services, car sharing and delivery robots. The collected data includes information about the mode in use, trip details, vehicle properties and detailed vehicle state information. In addition to that, the MDS standard specifies that both the operators and the municipality where the services are being offered, provide data to make sure that the mobility service is implemented in accordance with local rules and regulations, e.g. a *Geography* endpoints specification which is used to define boundaries, pick-up, drop-off or inaccessible zones or a *Policy* endpoints specification used to enforce compliance (e. g., number of vehicles deployed, speed limits).

Public transportation. This data is accessible directly from the providers, transportation organisations or from their system providers throughout Germany. A subset of this data is publicly accessible while the rest is restricted and only accessible on a contract base. For this project, real-time traveller information is available through the VDV 431 TRIAS-API [20]. The TRIAS-API comprises information covering multiple aspects of the transport network such as line schedules, departure monitors, individual routing or trip information. The process of querying information is usually initiated by an individual asking for certain information related to a trip.

Bicycles. During the survey period, 300 test subjects will be equipped with DASHBIKE[3] sensors. They are mounted on the seat post against the direction of travel at a 90° angle to the road. They are equipped with an HD camera, distance measurement via radar, GPS, fall detection with gyroscope and Wi-Fi and Bluetooth interfaces. As part of a crowdsensing campaign, test subjects record where in Bamberg motorists undercut the minimum distance of 1.5m when overtaking, accident locations and GPS trajectories.

Pedestrians. To monitor pedestrian movements and crowds, we installed Wi-Fi sensors in Bamberg's city center at places of tourist interest to detect visitors' Wi-Fi probe requests [2]. The probe request data and signal strength can be used to estimate the visitor frequency at the measurement points. This is a passive, non-intrusive and low-cost alternative compared to methods like camera-based or Lidar installations.

Parking. For our study, the main provider of parking data is Parking Pilot[4] followed by data obtained from parking garages equipped with digital parking systems[5]. We get coarse data collected from parking facilities which only monitor the total number of occupied parking spots at a given time as well as dense data collected from sensors monitoring each individual parking spot. The coarse data can be enriched by combining it with features extracted from the parking facilities entry and exit points, to determine the vehicle registration, type and duration for which each vehicle uses the parking facility for, and by additional information, such as parking rights and tariffs. The dense data can encompass

[2]https://zeusscooters.com/ Accessed on: 2023-09-15
[3]https://www.dashbike.de Accessed on: 2023-09-15
[4]http://parking-pilot.com/ Accessed on: 2023-09-15
[5]https://designa.com/ Accessed on: 2023-09-15

additional information specific to each parking spot. This can include parking spot user information, the status of the parking spot, the type of vehicle parked, i. e., based on the type of fuel used, emissions or size, the duration for which the parking spot is reserved or the location in the parking lot.

4 ANONYMIZATION TECHNIQUES

Privacy models such as k-anonymity are widely used in the literature, for instance when anonymizing network data [5]. However, anonymization techniques are rarely applied to mobility data in practice although that data compromises user privacy as it might be possible to generate individual mobility profiles [10, 17]. We believe that there is a limited number of practical adaption because anonymization of mobility data is more complex. For proper anonymization, it is – due to the nature of the data – often not sufficient to apply traditional anonymization techniques such as generalization, permutation, or hashing. In practice, mobility data is often anonymized by removing the start and end of trips and by using trip identifiers instead of user identifiers [10, 17]. Additionally, surveys [10, 17] have shown that some approaches use k-anonymity [1, 8], differential privacy [3, 13, 15], or dummy data [9, 11]. However, even if researchers use those anonymization techniques to protect users' privacy, someone might still infer sensitive information about the individuals. That is, not only the start and end of a trip represents sensitive information; there are also certain POIs along the route, which is often not considered when anonymizing mobility data. Additionally, the utility of the data might decrease if researchers use differential privacy or dummy data for the anonymization.

Previous Experiences: Those problems have already been identified by Psoido in a former project about anonymization with an optimized privacy-utility tradeoff. In that project, six anonymization techniques have been combined for an anonymization concept for mobility data. At first, there is an ID management, so that every trip is assigned a unique identifier, which is not related in any way to the users. As POIs might reveal information about users, they are then identified along the route of those trips. If there is any POI along the route, the trips are split at those points and the new trips are assigned unique identifiers. In the next step, the start and end of trips is removed.

To preserve the privacy of the users further, generalization is used for the GPS data. Coordinates are – depending on their initial accuracy – kept unchanged, rounded, or truncated at the end. Additionally, noise from the trips' probability distribution is added to existing trips, so that it is not possible to differentiate between real and synthetic data points. Further, this ensures that users can not be identified if there are any trips with few or very characteristic data points. In the last step, undersampling is used to account for GPS device specific patterns. That is, one device might record GPS data in the same location following a device specific distribution. To prevent user identification, all data points can be mapped to a specific distribution, e. g., a normal distribution. The anonymization concept was tested in practice with real bicycle data. For scooter or public transportation data, only synthetic data has been used so far.

Scientific approaches: The difficulties anonymizing mobility data are also discussed in scientific works. Jin et al. identified some of the most popular anonymization techniques for mobility data [10]. The authors selected the most cited works for each privacy model from recent publications. We rely on their findings and describe the identified approaches shortly to assess their practicability for the use case of intermodal mobility data in smart cities.

W4M [1] and GLOVE [8] both use k-anonymity as privacy model. To achieve that, the authors of W4M make us of the imprecision of location data sensors, e. g., sensors for GPS data. From a technical perspective, the approach basically groups trajectories into clusters so that at least k trajectories are within the radius of the sensors' imprecision. Thereby, the cluster membership is based on the EDR distance, which considers both space and time. Finally, to create

the anonymity set for the individual clusters, i. e., to make trajectories within a cluster similar enough, the authors apply spatio-temporal editing to the trajectories.

The idea of GLOVE is that most mobile traffic fingerprints in a dataset can be anonymized quite easily while keeping high data utility [8]. To make use of this property, the authors use specialized generalization potentially with suppression to achieve k-anonymity. That is, each sample is anonymized independently while minimizing the loss of data utility. To be precise, at first, the spatio-temporal accuracy loss is calculated for all fingerprints when merging two of them so that they are indistinguishable from each other. Then, the fingerprints with the least accuracy loss are merged until k-anonymity is reached.

The algorithm by Tu et al. [19] was designed to improve the privacy protection in comparison to previous works on k-anonymity as those approaches are vulnerable to semantic attacks, where an attacker uses POIs to break users' anonymity. Their approach achieves k-anonymity, l-diversity, and t-closeness. Hence, Jin et al. call it KLT in their survey [10]. KLT uses a similar procedure as GLOVE based on suppression and specific generalization by calculating the spatio-temporal resolution loss, merging the spatio-temporal data points, and trajectories [19]. The basic idea is that POIs are contained in trajectories in a region. For l-diversity, the approach ensures that there are enough categories of POIs in a region. For t-closeness, the distribution in one region and in the city must not differ more than a certain threshold, which is measured by the KL divergence in the paper.

DPT [9] provides privacy guarantees with differential privacy. In contrast to other privacy models, there are no assumptions about attacker knowledge. The idea of DPT is to use a generative model to publish synthetic data of trajectories. To preserve high utility, a novel sampling technique is used. The Laplace mechanism ensures differential privacy by using noisy counts in prefix trees, which are constructed based on a probabilistic model.

With DTPP [13] dummy data are created to protect individuals' privacy. The basic idea is that dummy trajectory data are created in a way that those data is similar to the real trajectory data. Further, the algorithm protects exposure locations of the individuals. That is, dummy trajectories are desinged so that attackers can not infer which routes are dummy data and which are real data based on the information that the route goes through the exposure location.

5 DISCUSSION AND CONCLUSION

Our investigation highlights that current anonymization techniques, which often concentrate on a single data type like GPS trajectories, do not adequately accommodate the diversity of data formats found in intermodal mobility chain analysis. Since our goal is to use explainable anonymization techniques, we disregard differential privacy approaches such as DPT in our implementation. For anonymizing the GPS data collected from e-scooters and public transportation the approaches W4M, KLT, and DTPP need to be evaluated. The GPS trajectories of cyclists collected by the DASHBIKE sensors are anonymized using an approach developed by Psoido. The pedestrian data is also already provided in an anonymized form [2]. The applicability of the W4M, KLT, and DTPP approaches to parking data needs to be evaluated. Selecting the appropriate anonymization method requires clear criteria based on specific use cases, which we need to develop with stakeholders and test for acceptance.

As our research brings together data from multiple sources with varying levels of granularity and anonymization, we plan to develop probabilistic methods for identifying transitions between mobility types to provide a holistic view of mobility patterns. While GLOVE concerns the anonymization of mobile traffic data, the generalization approach is worth evaluating in terms of its applicability to intermodal mobility data.

Based on Psoido's research, we propose to combine multiple techniques to address the challenges of anonymizing intermodal mobility data from multiple sources. This approach mitigates the weaknesses of each method and also

improves comprehensibility by selecting explainable methods, making anonymization more accessible. Unlike the Psoido study, which used bicycle and synthetic data, we use multiple real-world data sources that provide practical insights into anonymization and actual mobility patterns. Real-world context reveals nuances that may be missing in controlled environments.

In summary, *explanym* contributes to the evolving field of mobility data anonymization by combining theoretical approaches, learning from real-world experiences, and connecting disparate data sources.

ACKNOWLEDGMENTS

The project *explanym* - Explainable Anonymization of Intermodal Mobility Data is funded by the German Federal Ministry of Education and Research (BMBF).

REFERENCES

[1] Osman Abul, Francesco Bonchi, and Mirco Nanni. 2010. Anonymization of moving objects databases by clustering and perturbation. *Inf. Syst.* 35, 8 (2010), 884–910.

[2] Leonie Ackermann, Christoph Baum, Syed Ibrahim Khalil, Aleksandr Litvin, and Daniela Nicklas. 2023. Privacy-aware Publication of Wi-Fi Sensor Data for Crowd Monitoring and Tourism Analytics. *Submitted to 1st ACM SIGSPATIAL International Workshop on Geo-Privacy and Data Utility for Smart Societies* (2023).

[3] Miguel E. Andrés, Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: differential privacy for location-based systems. In *2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13*. ACM, 901–914.

[4] BBSR. 2017. Smart City Charter - Making digital Transformation at the Local Level Sustainable. *Scientific Support: Federal Institute for Building, Urban Affairs and Spatial Development (BBSR), Division I* (2017). https://www.bbsr.bund.de/BBSR/DE/veroeffentlichungen/sonderveroeffentlichungen/2017/smart-city-charta-de-eng.html;jsessionid=940547B45D1FE52CA3F5B11C67EA175C.live11293

[5] Niels Van Dijkhuizen and Jeroen van der Ham. 2018. A Survey of Network Traffic Anonymisation Techniques and Implementations. *ACM Comput. Surv.* 51, 3 (2018), 52:1–52:27.

[6] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407.

[7] GBFS. 2023. *General Bikeshare Feed Specification (GBFS) Specification Reference.* https://gbfs.org/specification/reference/ Accessed on: 2023-09-15.

[8] Marco Gramaglia and Marco Fiore. 2015. Hiding mobile traffic fingerprints with GLOVE. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies, CoNEXT 2015*. ACM, 26:1–26:13.

[9] Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M. Procopiuc, and Divesh Srivastava. 2015. DPT: Differentially Private Trajectory Synthesis Using Hierarchical Reference Systems. *Proc. VLDB Endow.* 8, 11 (2015), 1154–1165.

[10] Fengmei Jin, Wen Hua, Matteo Francia, Pingfu Chao, Maria E. Orlowska, and Xiaofang Zhou. 2023. A Survey and Experimental Study on Privacy-Preserving Trajectory Data Publishing. *IEEE Trans. Knowl. Data Eng.* 35, 6 (2023), 5577–5596.

[11] Ryo Kato, Mayu Iwata, Takahiro Hara, Akiyoshi Suzuki, Xing Xie, Yuki Arase, and Shojiro Nishio. 2012. A dummy-based anonymization method based on user trajectory with pauses. In *International Conference on Advances in Geographic Information Systems, SIGSPATIAL'12*. ACM, 249–258.

[12] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *Proceedings of the 23rd International Conference on Data Engineering*. IEEE Computer Society, 106–115.

[13] Xiangyu Liu, Jinmei Chen, Xiufeng Xia, Chuanyu Zong, Rui Zhu, and Jiajia Li. 2019. Dummy-Based Trajectory Privacy Protection Against Exposure Location Attacks. In *Web Information Systems and Applications - 16th International Conference, WISA (LNCS, Vol. 11817)*. Springer, 368–381.

[14] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. *L*-diversity: Privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1 (2007), 3.

[15] Àlex Miranda-Pascual, Patricia Guerra-Balboa, Javier Parra-Arnau, Jordi Forné, and Thorsten Strufe. 2023. SoK: Differentially Private Publication of Trajectory Data. *Proc. Priv. Enhancing Technol.* 2023, 2 (2023), 496–516.

[16] Open Mobility Foundation. 2023. *Release 2.0.0.* https://github.com/openmobilityfoundation/governance/wiki/Release-2.0.0 Accessed on: 2023-09-15.

[17] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. 2019. The Long Road to Computational Location Privacy: A Survey. *IEEE Communications Surveys & Tutorials* 21, 3 (2019), 2772–2793.

[18] Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 10, 5 (2002), 557–570.

[19] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin. 2019. Protecting Trajectory From Semantic Attack Considering k-Anonymity, l-Diversity, and t-Closeness. *IEEE Trans. Netw. Serv. Manag.* 16, 1 (2019), 264–278.

[20] Verband Deutscher Verkehrsunternehmen. 2014. *Interface specifications of VDV 431-1/-2.* https://www.vdv.de/projekt-ip-kom-oev-ekap.aspx 2023.