Alexandra Kapp, Saskia Nuñez von Voigt, Helena Mihaljević, Florian Tschorsch

Towards mobility reports with user-level privacy

Towards Mobility Reports with User-Level Privacy

Alexandra Kapp[1] Saskia Nuñez von Voigt[2] Helena Mihaljević[3]
Florian Tschorsch[4]

[1]ORCiD: 0000-0002-8348-8958, Hochschule für Technik und Wirtschaft Berlin, University of Applied Sciences,
kapp@htw-berlin.de
[2] Distributed Security Infrastructures Technische Universität
[3]ORCiD: 0000-0003-0782-5382, Hochschule für Technik und Wirtschaft Berlin, University of Applied Sciences
[4]ORCiD: 0000-0001-6716-7225, Distributed Security Infrastructures Technische Universität Berlin

Abstract

The importance of human mobility analyses is growing in both research and practice, especially as applications for urban planning and mobility rely on them. Aggregate statistics and visualizations play an essential role as building blocks of data explorations and summary reports, the latter being increasingly released to third parties such as municipal administrations or in the context of citizen participation. However, such explorations already pose a threat to privacy as they reveal potentially sensitive location information, and thus should not be shared without further privacy measures.

There is a substantial gap between state-of-the-art research on privacy methods and their utilization in practice. We thus conceptualize a mobility report with differential privacy guarantees and implement it as open-source software to enable a privacy-preserving exploration of key aspects of mobility data in an easily accessible way. Moreover, we evaluate the benefits of limiting user contributions using three data sets relevant to research and practice. Our results show that even a strong limit on user contribution alters the original geospatial distribution only within a comparatively small range, while significantly reducing the error introduced by adding noise to achieve privacy guarantees.

Keywords: human mobility data, differential privacy, user-level privacy, exploratory data analysis, mobility report

1 Introduction

Mobility data is becoming increasingly important - for urban planning [1], traffic management [2], and smart city applications [3]. To tackle challenges posed by the climate crisis, cities require movement data to address smart traffic management [4, 5], shift car use to climate-friendlier options [6], offer broader access to public transport [7], and steer ever-growing mobility options such as e-scooters [3], to name a few. The current pandemic further highlights the usefulness of such data as these form the basis for analyses [8, 9] or simulations of pandemic progression [10].

At the same time, mobility data comprises high-dimensional, sensitive information and thus poses challenges towards ensuring privacy. For example, just a few recorded locations are sufficient to re-identify an individual [11], potentially revealing private information such as home and work locations, political interests or religious beliefs. Recent discussions on storage and processing of mobility data for pandemic management highlight that many citizens are willing to disclose their movement data only under extensive privacy guarantees [12, 13]. Uncertainty about how to adequately protect user privacy and thus comply with applicable data protection laws, for example, makes many mobility providers reluctant to share their usage data with cities or otherwise release it.

Transparency of policy making can therefore only be guaranteed to a limited extent. The Open Mobility Foundation facilitates data exchange between mobility providers and cities, thereby also striving to foster transparency of policy processes to the public. At the same time, they address the inherent privacy issues of the release of raw mobility data and suggest data minimization, aggregations and reports as potential countermeasures [14]. For a lot of decisions reporting aggregate statistics is sufficient as many use cases need no fine granular trajectory data. In Germany, for example, according to road traffic regulations a street can be redesignated as a bicycle street if a high volume of bicycle traffic can be proven. Similarly, aggregated mobility data is sufficient to monitor the impact of measures like new bike infrastructure or shared-mobility docking stations on usage behavior. Beyond monitoring tasks, aggregated statistics are highly valuable for more complex use cases, such as traffic models [15] or identification of optimal e-scooter parking zones [16], since these require exploratory data analyses as a first step to familiarize with the data and evaluate its suitability for the aspired use case. While aggregations provide some privacy they do not come with privacy *guarantees*; in fact there are successful attacks on aggregated mobility data [17] that recover entire trajectories of individuals.

We propose a mobility report comprising most relevant and frequently applied mobility measures that is equipped with privacy guarantees. For this purpose, we draw on extensive research on the analysis of movement data that requires the simultaneous examination of spatial and temporal properties together with the moving object itself [18]. Depending on the data source, mobility data can differ greatly in its format and structure, spatial and temporal granularity, or content. Thus, it is hardly possible to comprise a mobility report that suits all data sources and use cases. We carefully aimed to maintain the report as general as possible while knowing that it cannot fit all needs. Specifically, our report provides high level key insights to mobility data sets that could be used to provide information to public administrations, to the public, or for company's internal knowledge sharing. The report is especially suited for data containing single trips with an origin and destination (e.g., bike-sharing transaction history) and use cases about mobility on an aggregated level (e.g., usage of a new introduced bike-sharing service). It is also aimed at the analysis of staypoints and does not cover route information such as traffic volumes or the average speed on route segments. Since mobility data includes sensitive information, we additionally see the need for privacy protection and apply methods for securing differential privacy [19] which is considered future-proof, e.g. the privacy guarantees hold regardless of the attacker knowledge or additional information (that becomes available later).

Unlike an iterative approach of arbitrary exploratory analyses, a report with predefined measures allows for optimizing the allocation of a predefined privacy budget, as exploratory analyses might otherwise come at a high privacy cost [20]. In order to keep the privacy costs as low as possible while ensuring high usefulness of the report, both the selection and detailed design of the analyses were optimized.

While differential privacy is well understood for aggregations of general tabular data, there is little research on how to apply such a mechanism to aggregations of mobility data. Due to the nature of mobility data, differential privacy is difficult to apply as mobility traces of people are inherently individual with arbitrarily long sequences and spatial and temporal dimensions covering an infinite range of values. Broadly speaking, differential privacy guarantees that an adversary can only identify the membership of an individual within a data set at a defined probability level. The lower the identification probability shall be maintained the more noise has to be added. Generally, the smaller the number of buckets of aggregations the lower the share of added noise: consider adding noise of +/- 5 items to one bucket with a count of 100 or to 10 buckets each with a count of 10. In the first case, the noisy number of items will differ by at most 5%, while in the second case, the difference per bucket will be up to 50%. The number of buckets of mobility data can be adjusted at the spatial dimension (e.g., aggregation on a 100 m grid vs. 1 km grid) and the temporal dimension (e.g., hourly aggregation vs. weekly aggregation). Also, the space of possible sequences can be adjusted, e.g., by only considering start and end connections instead of entire routes. However, the number of buckets increases drastically once certain attributes are cross tabulated, for example, looking at spatial distributions per hour of day, or investigating the relation of origins and destinations. Data sparsity is an additional issue, as the records are usually not equally distributed over all buckets. For example, a city center usually has high visit counts while other areas further outside only get a few visits. Thus, differentially private values are only reliable for certain buckets while others mainly consist of noise.

We aim to gain insights how differential privacy can be applied to aggregations of mobility data and what level of utility could be expected when applied to typical, real-world data sets. We therefore evaluate key measures of the report under different privacy scenarios on three data sets that are used extensively, especially in mobility research, e.g., to infer transportation modes [21], mine locations of interest [22], or to investigate the influence of parking-regulations on car-ownership [23]. This provides an important orientation for the trade-off between privacy and utility of real mobility data sets. Moreover, a typical mobility data set can contain an arbitrary number of records per user; an upper bound of user contribution is typically not set or known upfront. Intuitively, the more records a user contributes to a data set the more likely they will be identified. Thus, to provide differential privacy at a user level, more noise needs to be added according to the potential contribution of a single user. In order to achieve acceptable estimates of the sensitivity of the respective aggregation functions when applying noise to ensure user-level privacy, it is useful to limit the records per user [24, 25, 26, 27]. However, it is not clear how to choose such an upper bound. We thus analyze the trade-off between privacy and utility under user-level privacy guarantees and different choices of the limit of user contribution.

A report for mobility data with differential privacy guarantees has the potential to simplify and accelerate secure analysis and releases of movement data. Analyses of mobility data require specific geospatial skill sets and resources. With regard to privacy it should be noted that large companies such as Google, Apple, or Microsoft already make use of differential privacy [28], while smaller companies and the public sector often lack knowledge and resources required to keep up with big tech players [29]. To facilitate access and increase usability, we thus provide an implementation of our differentially private mobility report as open-source Python code.

Our Contributions. After discussing research on mobility data analyses and appropriate measures for guaranteeing differential privacy in Section 2, we identify elementary measures commonly applied to human mobility data (Section 3) and compile a mobility report. We provide a differentially private version of the report and explain the resulting implications (Section 4). We analyze the effect of user-level privacy and the interplay with different bounds of user contribution on three data sets (Section 5), and provide practical implications of a user-level differentially private mobility report (Section 6). In this way, we support data analysts in understanding the implications of a differentially private mobility report.

2 Related Work

Data analysts use exploratory data analyses (EDA) to become familiar with data sets, usually focusing on visual inspection. They aim to understand the data, check its validity, detect outliers, identify possible patterns and assess its suitability for further analyses or models [30, 31]. The concept of EDA was established by Tukey, who, for example, promoted the use of the five-number summary to describe the distribution of numerical data [32]. The issue of predefining a set of measures or analyses for an EDA has already been addressed in literature, e.g., [33]. In addition, implementations of standardized data explorations have recently been made available, e.g., in form of the Python package `pandas_profiling` [34], making exploratory analyses accessible for a broader group of practitioners.

The analysis of movement data poses additional challenges, as it requires the simultaneous examination of spatial and temporal properties together with the moving object itself [18]. Therefore, mobility data need specific analyses in addition to standard EDA methods that can capture and convey their information. Andrienko et al. [35] have massively contributed to visual analytics of different types of movement data that can be considered valuable for the creation of a mobility data report. However, no attempt has been made to predefine a set of analyses for a particular type of movement data and analytical context such as human mobility in cities. Graser [36] argues that a standardized protocol for an EDA can hardly cover all kinds of movement data at once and presents a protocol for an EDA with the goal to identify problems in GPS data, such as unrealistic jumps in individual trajectories. The Python package `scikit-mobility` [37] implements a comprehensive collection of measures suitable for human movement data, without combining a subset of them into a single report. Additionally, the package includes methods for privacy risk assessments; however, no privacy mechanisms are provided that can be applied to the computation of statistics.

In the context of publishing privacy-preserving aggregated statistics, differential privacy has been proposed for different cases, focusing on relational data and the interactive setting, i.e., data can only be accessed via certain database queries [38, 39, 40]. For mobility data a multitude of approaches address differential privacy in the non-interactive setting, i.e., publishing synthetic mobility data [41, 42, 43, 44]. However, these works assume that a user does not contribute multiple items. Therefore, we focus on releasing mobility measures to understand the implications of user-level and item-level privacy.

For mobility data a multitude of approaches address differential privacy in the non-interactive setting, i.e., the publication of anonymized historical data in their raw format by, e.g., generating synthetic data [42, 45, 44, 46]. Such data releases are commonly motivated by the lack of openly available mobility data for a wide range of (not yet fully specified) use cases. However, often the needed analyses are known in advance which allows the application of privacy-preserving techniques which are more precise and targeted for the respective use case.

Additionally, many existing approaches assume that a user does not contribute multiple items. Liu et al. [26] studied the utility privacy trade-off for user-level privacy and compared it to the item-level counterpart. While they evaluated it in the context of learning discrete distributions, we focus on a multitude of aggregations comprising a mobility report. Wilson et al. [27] present an approach to user-level privacy for aggregations by bounding user contribution. They account for situations in which the categories used for grouping the data are not known a-priori. However, they guarantee approximate differential privacy. Aktay et al. [9] follow their overall approach to compute aggregated mobility metrics in the context of the COVID-19 pandemic. They account for user-level differential privacy by bounding user contribution to four records. We adapt their differential privacy mechanism, mainly consisting of adding noise to aggregated statistics and bounding user contribution, and extend it to a variety of mobility measures and evaluate the impact of the chosen bound for user contribution.

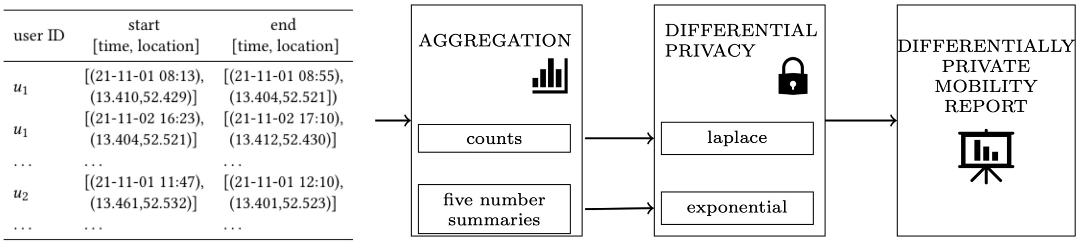
Amin et al. [24] theoretically analyze the bias-variance trade-off of bounding user contribution. Allowing users to contribute large amounts of data may add excessive noise to protect a few outliers, while limiting users to small contributions keeps noise levels low at the cost of potentially discarding relevant information and thus introducing bias. Their results show that an optimal contribution limit can be found for which the expected error of differentially private empirical risk is minimal, but that there is no contribution bound that is sufficient to eliminate both bias and variance, even in the limit case of infinite data [24]. In this paper, we make similar considerations and evaluate the influence of bounding user contribution on suitable error measures. Additionally, we discuss relevant aspects that determine the trade-off between item-level and user-level differential privacy and show consequences of a differentially private mobility report.

3 Mobility Report

Following the idea of a report with a predefined set of measures for general tabular data, as implemented by the `pandas_profiling` package, we propose a urban human mobility data report that includes statistics commonly applied for the exploration of mobility data sets reflecting movement of individuals in an urban area [47, 48, 37, 49, 50]. This will serve as a basis for practitioners to gain fundamental insights from human movement data and establish a basis for further decisions.

3.1 Mobility data

Human movement analyses usually focus on stay-points where a certain amount of time is spent, e.g., 'home', 'work' or 'supermarket' [47], in contrast to way-points that are typically only passed by during a trip. In terms of privacy risks, the stay-points are also of special interest: although there are attacks that use mobility features such as speed, direction, or distance [51], most attacks focus on important locations [52]. We will thus restrict to a generalized form of human mobility defined as a collection of single trips comprised of origin-destination pairs, in particular, we will ignore way-points. This results in the following requirements for input data that are satisfied by many common urban mobility data sets like surveys, routing queries, public transport smart card check-ins and -outs, or shared mobility rentals.



user ID	start [time, location]	end [time, location]
u_1	[(21-11-01 08:13), (13.410,52.429)]	[(21-11-01 08:55), (13.404,52.521])
u_1	[(21-11-02 16:23), (13.404,52.521)]	[(21-11-02 17:10), (13.412,52.430)]
...	...	...
u_2	[(21-11-01 11:47), (13.461,52.532)]	[(21-11-01 12:10), (13.401,52.523)]
...	...	...

Figure 1: Overview of our process for a differentially private mobility data report.

Let U be a set of user identifiers and $P = \{p = (t, l)\}$ a set of spatio-temporal points p consisting of a timestamp t and a location l referring to a given coordinate reference system (CRS), e.g., geographical latitude and longitude. A record of the input data set T consists of a user identifier $u \in U$ and two spatio-temporal points, the start and end point, that together define a trip tr. We show a mobility data set example in Figure 1. Origins and destinations are mapped onto a tessellation. This is needed to spatially aggregate the data for statistical analyses, modeling tasks [49], and comprehensive visualizations of trips that are not cluttered with lines [53]. Typical tessellations are administrative boundaries or evenly distributed grids [49]. We assume that such a tessellation is provided for the report.

3.2 Mobility measures

Human mobility can mainly be analyzed from three perspectives: places that are visited, movements between places, and mobility characteristics of individuals [54]. For *place analyses* the number of movements originating or arriving at a location are of interest, while *trip analyses* focus on where people are coming from or going to. The analyses of mobility characteristics of individuals, which we refer to as *user analyses*, puts a single person into the center of attention. All movements of a person are examined together, for example to evaluate the spatial range individuals typically move in. We structure the proposed report in accordance with the described analysis perspectives in addition to a general group of *overview analyses.* The first three columns of Table 1 provide an overview of all statistics included, which are explained in detail below.

The first group of analyses gives an overview of basic information, including the number of *trips*, *users* and *locations.* Temporal properties are examined in three different ways: trip counts are aggregated to the number of *trips over time*, where the time interval is chosen automatically to be a day, week or month depending on the time range covered by the data. This measure reveals the temporal coverage and completeness of the data, the potential existence of (un)intentional interruptions and patterns of seasonality. It is provided as trip counts per time interval and as a five-number summary. Weekly and daily cycles, which capture patterns that are characteristic for human mobility [55, 48], can be inferred from the number of *trips per weekday* and *trips per hour* disaggregated by weekday and weekend, respectively.

The second group, place analysis, provides information on the geospatial distribution of the data in terms of the provided tessellation. The main measure computes the total number of *visits per location* over the entire time period of observation. As points of trips can potentially lie outside the tessellation, the number of outliers is provided as well. Note that there are twice as many visits as trips, since both start and end point are considered.

The frequency of location visits usually differs depending on the time of day [48]. Therefore, spatial distributions are also disaggregated as *visits per destination by time* and split by weekday and weekend. To limit the number of outputs, we use a default time span of four hours to create six time windows unlike, e.g., hourly time windows that would result in 24 views. The default time windows, which are adjustable in the package implementation, are defined as follows: 2:00 - 6:00, 6:00 - 10:00, 10:00 - 14:00, 14:00 - 18:00, 18:00 - 22:00, 22:00 - 02:00. Only destinations are considered to correctly represent spatio-temporal patterns. For illustration, consider the following example: in the morning people commute from the outskirts to the city center and vice versa in the evening. If both origins and destinations would be considered, the outskirts and the city center would be highly frequented in the morning and evening, thus not revealing important temporal patterns.

The third group covers trip analyses. Trips are aggregated and counted according to origin and destination tiles, resulting in an OD matrix, named *OD flows.* Additionally, the *travel time* and *jump length* [49] are computed, the latter representing the geographical straight-line distance between the

Table 1: Overview of mobility measures and sensitivity.

Group	Measure	Function	Sensitivity
Overview	trips	count_{tr}	M
	users	count_{u}	1
	locations	count_{p}	$2 \cdot M$
	trips over time	counts_{tr},five-number summary$_{tr}$	M, M
	trips per weekday	counts_{tr}	M
	trips per hour	counts_{tr}	M
Place analysis	visits per location	counts_{p}	$2 \cdot M$
	visits per destination and time	counts_{tr}	M
Trip analysis	OD flows	counts_{tr}	M
	travel time	counts_{tr}, five-number summary$_{tr}$	M, M
	jump length	counts_{tr}, five-number summary$_{tr}$	M, M
User analysis	trips per user	counts_{u}, five-number summary$_{u}$	$1, 1$
	radius of gyration	counts_{u}, five-number summary$_{u}$	$1, 1$
	locations per user	counts_{u}, five-number summary$_{u}$	$1, 1$
	mobility entropy	counts_{u}, five-number summary$_{u}$	$1, 1$
	time between trips	counts_{tr}, five-number summary$_{tr}$	M,M

trip's origin and destination. They are reported in form of five-number summaries and histograms, i.e., counts. Travel time and jump length typically decay as a power law [56] with many short times and distances and few very long trips. An unbound histogram would therefore be highly right-skewed, either with many small bins or a few non-expressive bins at the lower end of the histogram. To limit the number of histogram bins for a more comprehensive representation, we propose to cut them off at a defined maximum based on domain knowledge and to additionally provide the number of outliers above the defined threshold. In addition to usability reasons, a pre-defined maximum helps to control the privacy-utility trade-off as will be shown in Section 4.

The last group on user analyses compiles the following measures, summarized as histograms and five-number summaries. The number of *trips per user* computes a user's overall contribution to the data set. *Locations per user* represents the distinct locations visited by a user [48] and describes the diversity of locations a user visits. The *mobility entropy* entails related information: it is defined as the Shannon entropy of the user's visits which quantifies the probability of predicting a user's whereabouts [57, 48, 49]. The *radius of gyration* is the characteristic distance traveled by a user [58, 48, 49], computing the spread of all locations visited by an individual around their center of mass. Again, the distribution of radii of gyration follows a power law, thus setting cut off values for the histogram is beneficial for usability and privacy reasons. Further entailed is the *time between trips* which computes the amount of time between the end time of a user's trip and the beginning of the next one. It thus quantifies the temporal density of user trips in the data, i.e., if user movements are recorded every other hour, day, week or month.

4 Differentially Private Mobility Report

4.1 System Model and Design

Before describing our system for a differentially private mobility data report, we introduce respective definitions and notations used throughout this paper and illustrate the intermediate steps to such a report as depicted in Figure 1. Assume a company owns a mobility data set T and needs to release a report. For a mobility data report, statistics defined in Section 3 are computed and visualized. Despite the aggregation of data, the mobility report consists of summary statistics which are vulnerable to reconstruction attacks [40]. By observing answers from measures, such as listed in Table 1, an attacker can recover secret information, such as trips.

Differential privacy provides mathematical guarantees for the privacy of an individual [19]. The concept of differential privacy is that the output of an algorithm $\mathcal{A}$ remains nearly unchanged if the records of one individual are removed or added. In this way, differential privacy limits the impact of a single individual on the analysis outcome, preventing the reconstruction of an individual's data.

Definition 1 (Differential Privacy)**.** *Let Range($\mathcal{A}$) be a randomized algorithm that takes a mobility data set T as input and outputs a value from some output space Range($\mathcal{A}$). For an $\varepsilon > 0$, $\mathcal{A}$ is said to be ε-differentially private, if for all pairs of data sets T_1 and T_2 differing in all records of an arbitrary but fixed user, and all outputs $O \subseteq Range(\mathcal{A})$,*

$$P[\mathcal{A}(T_1) \in O] \le \mathrm{e}^{\varepsilon} \cdot P[\mathcal{A}(T_2) \in O].$$

The parameter ε captures the privacy loss and determines how similar the randomized outputs are based on T_1 and T_2, and thus specifies the impact of a single individual's data records on the output.

In differential privacy literature, the definition usually considers data sets differing in one record, assuming that every user makes exactly one contribution. This is for instance the case when each record of a data set corresponds to a complete trajectory of a single user. However, typical mobility data sets in practice do not satisfy this assumption, containing multiple records per user. In this case, the classical definition of differential privacy would protect an item, such as a single trip. However, our used definition of differential privacy protects the privacy of a user. In the remainder of the paper, we will thus distinguish between *item-level privacy* and *user-level privacy* [19].

An important notion in this context is the sensitivity of a function f which corresponds to the maximum difference that an output can change by removing or adding a record (item-level) or all records of a user (user-level) [19].

Definition 2 (Sensitivity)**.** *Let T_1 and T_2 be two data sets differing in one record (item-level) or all records of a user (user-level), respectively. The L_1-sensitivity of f is defined as $\Delta f = \max_{(T_1,T_2)} |f(T_1) - f(T_2)|_1$ for any such T_1 and T_2.*

In our mobility report, most of the functions, listed in Table 1, are counts and as such output numeric values. For example, the function *count* outputs a single number, while *counts* outputs a number for each category and bin, respectively. A common mechanism for numeric functions is the Laplace mechanism, where calibrated noise is added to the function's output, drawn from a Laplace distribution *Lap*() [19].

Definition 3 (Laplace mechanism)**.** *Let $f : T^n \to \mathbb{R}^k$ with arbitrary domain T. The Laplace mechanism is defined as $\mathcal{A}(T) = f(T) + (Y_1, \ldots, Y_k)$ where Y_i is a random variable drawn from $Lap\,(\Delta f/\varepsilon)$ with mean 0 and variance $2(\Delta f/\varepsilon)^2$.*

Note that for *counts*, k equals the number of categories/bins, e.g., locations, while $k = 1$ for the *count* function. The magnitude of noise is calibrated according to the sensitivity Δf of a function.

If we assume item-level privacy for our report, this means for the number of *trips over time* that noise with $\Delta f = 1$ is added to each count, since removing or adding one trip changes the count by one. While this effectively hides the presence of one trip of an individual, it does not protect the presence/absence of an individual with multiple trips. In other words, removing or adding all trips of a user changes the number of *trips over time* by the amount of trips of the corresponding user instead of just one.

Suppose we are interested in the number of *OD flows.* A user has made 10 trips. If we remove the user, the counts will always change by 10: If the user made all the trips between the same origin and destination, the number for this OD flow changes by 10. The other counts remain unchanged. If the user made the 10 trips between varying OD pairs, the counts for each respective of pair change by 1. This means that the maximum influence of a user on such a measure f and therefore its sensitivity $\Delta f = 10$. In practice, however, a user can contribute an arbitrary number of trips and the sensitivity of f is thus unbounded.

4.2 Bounded user contribution

In this section, we present the differentially private functions of our mobility measures proposed in Section 3. We list the functions and sensitivity to guarantee user-level privacy in Table 1.

In order to limit the sensitivity, we need to limit the number of possible trips M of a user. If we choose the highest number of trips a user has in our data set for M, we assume local sensitivity. Local sensitivity depends on the data set and not only on the function. Therefore, the local sensitivity or the maximum number of trips may jeopardize the privacy of a user.

Sampling bounds the number of trips of a single user to M and removes the remaining records. The places people visit is highly predictable for most individuals [58, 57]. There are only few places that a person visits regularly, usually the home and work place, which make up the majority of locations in a person's mobility pattern [59]. Therefore, we assume that the global geospatial patterns of a mobility data set remain intact even though only a small sample of each person's trips are included.

The question arises as to what maximum the number of trips should optimally be set at. For example, Aktay et al. [9] have limited the number to 4 in their differentially private mobility report, without explaining this further. We will look at this question in detail as part of our evaluation.

4.2.1 Counts

The count functions based on users ($count_u$) have a sensitivity of 1, since removing a user, no matter how many trips they make, only changes one count by 1. E.g., the *trips per user* are represented with a histogram, each bin representing a possible number of trips. Removing/adding a user with their trips will only change the count for one bin by 1. The count functions based on trips (count_{tr}) have a sensitivity of M as explained in the previous example. Since a trip consists of two points (start and end), the count functions based on points (count_p) have a sensitivity of $2 \cdot M$.

To this end, we guarantee differentially private counts. However, e.g., visited locations or reported categories can reveal the identity of an individual. For example, if only tiles that were actually visited are included in the report, we can infer that all reported tiles were visited at least once, while all tiles that were not included are not part of the data set. Thus, we consider all tiles within the tessellation as given to obtain a finite number of geometric shapes and apply the Laplace mechanism to each of the tiles, including those with a count of zero. Therefore, there is no certainty which tiles have actually been visited and which have not. All points outside the provided tessellation are summarized as a single noisy count of outliers.

The tessellation defines categories for the spatial dimension. Categories for temporal dimensions, which is reflected by *trips over time*, are aggregated to specific time intervals, such as day, week or month. Empty intervals in between should be filled with 0 values so that noise can be applied as well. While the tessellation also provides a fixed bounding of the spatial extent, there is another issue for *trips over time*: revealing the exact first and last day of the entire time interval violates the privacy. The same issue of leaking minimum and maximum values applies to any count-based measure that has no pre-defined categories or bins, namely, *travel time*, *jump length*, *trips per user*, *time between trips*, *radius of gyration*, *locations per user* and *mobility entropy*. Similar to the given tessellation we can cut off bins at a defined minimum and maximum based on domain knowledge. Data points outside this interval are summarized as a single noisy outliers count. Instead, we can also determine the minimum and maximum differentially private to obtain bounds. Since the minimum and maximum are included in the five-number summary we do not need to compute them twice.

4.2.2 Five-Number Summary

The five-number summary can be returned differentially private with the exponential mechanism [60]. This mechanism does not add noise to the output. Instead it returns the best answer from a set based on a scoring function. Differential privacy is guaranteed since sometimes an output is returned even though it has not the highest score.

Definition 4 (Exponential mechanism)**.** *Given an input data set T the Exponential mechanism [60] randomly samples an output $O \subseteq Range(\mathcal{A})$ with a probability proportional to $e^{\frac{\varepsilon s(T,O)}{2\Delta s}}$, where s is the scoring function and Δs the corresponding sensitivity.*

We use the exponential mechanism to determine the five-number summary including the minimum and maximum. In this case the scoring function is a rank function of the sorted input. Since we determine the index of an element, a user with M trips influences the output by M. Therefore, the sensitivity for the five-number summary in Table 1 is the same as that for counts. Note that the element returned by the exponential mechanism is always a member of the set $\text{Range}(\mathcal{A})$. This is reasonable for a finite set where a noisy response is not useful. Therefore, a pre-defined minimum/maximum is more privacy-preserving and should be preferred as far as possible.

5 Application and Evaluation

Table 2: Overview of evaluation data sets

			Tessellation			Trips per user			
	Trip count	User count	Tile count	Tessellation area	Mean tile area	Max	Median	Mean	SD
GEOLIFE	18,670	182	962	3,566 km^2	3.7 km^2	2,153	27.5	102.6	250.0
MADRID	222,744	75,208	1,259	8,031 km^2	6.4 km^2	20	2.0	3.0	1.5
BERLIN	1,417,134	378,759	386	891 km^2	2.3 km^2	16	4.0	3.7	1.8

5.1 Implementation and Provision of Code

The differentially private mobility data report proposed in Section 4 is implemented as a python code package[1] and provided as open source code under the MIT license. The implementation is inspired by the pandas_profiling package. It expects a *pandas DataFrame* [61] in the required format as input and outputs a report as an HTML file. This serves as an initial step to provide an easy to use implementation of the report in practice. We envision to further develop the package to comply with common standards and to adjust functionality based on testings with practitioners.

5.2 Experimental Setup

In the following, we compare the deviation of the mobility report from differentially private variants. For this purpose, we focus on four selected mobility measures, introduced in Section 3, that can be considered typical and unique for mobility data: *number of trips*, *visits per location*, *OD flows* and *radius of gyration.* For a full analysis of error measures for the entire report we refer the reader to our online repository[2].

Below, we describe the data sets, the choice of parameters for the privacy guarantees, and the error measures to quantify the deviations.

5.2.1 Data sets

We use three data sets, denoted by MADRID, GEOLIFE and BERLIN, commonly used in research and applications on human mobility. The first two are real, open data sets; BERLIN is a synthetic data set created with a traffic simulation software.

The data set MADRID [62, 63] is a survey on the mobility behavior of 75,208 residents of the Community of Madrid. Participants were asked about each trip they made on one weekday (Monday to Thursday) they selected between February and May 2018. The data set contains a tessellation consisting of 1,259 irregular sized traffic cells [64] that serve as the possible origins and destinations of the survey. 2,007 points were outside the given tessellation. As our input specification requires coordinates for origins and destinations, we replace the tile IDs with the coordinates of the respective tile centroids.

GEOLIFE was collected and released as part of the eponymous project [21, 22, 65] in which movements of 182 participants were recorded between 2007 and 2012 using GPS devices. Users were not continuously tracked over the entire time period; instead, single trips were recorded when users actively started tracking. Therefore, the time period and number of recorded trips differ greatly between individual users. Nine users contribute almost half of the trips with more than 400 trips each, while half of all users recorded only 27 trips or less. As we only use start and end locations, all way-points of the trips are removed for our purposes. Most traces are located in Beijing, China. As no tessellation is provided, we use a hexagonal grid based on the H3 geospatial indexing system[3] that covers the majority of data points within the Beijing center with an H3 resolution of 7, omitting 3,583 (9.5 %) of the records.

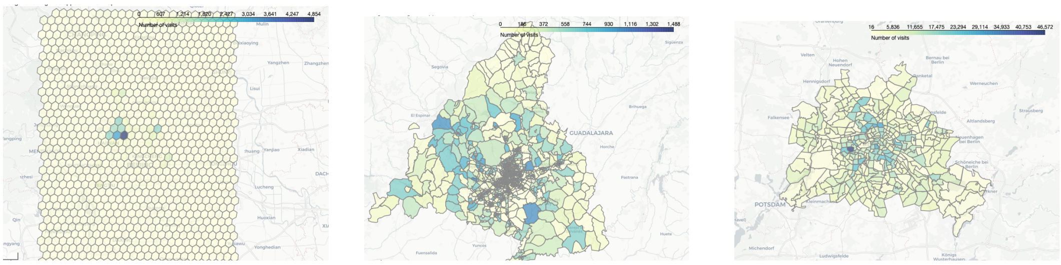
The BERLIN data set is produced by open-source traffic simulation software TAPAS [66] of German Aerospace Center and has been calibrated to the mobility behavior of the population of Berlin on a typical workday based on a survey in 2017. We work with a 10 % sample of the entire Berlin population which results in 378,759 users and 1,35 M trips and use traffic cells of Berlin [67] as a tessellation.

In Table 2 we show the main characteristics of the selected data sets and in Figure 2 we visualize their tessellations and the mobility measure of *visits per location.* Particularly relevant for consideration of user contribution bounds is the difference in user contributions per data set: while the distribution of trips per user in GEOLIFE is strongly skewed to the right, those of MADRID and even more BERLIN are

[1] Link to the package repository: https://github.com/FreeMoveProject/dp_mobility_report/

[2] Link to evaluation repository: https://github.com/FreeMoveProject/evaluation_dp_mobility_report

[3] https://h3geo.org/



(a) GEOLIFE (b) MADRID (Tessellation by [64]) (c) BERLIN (Tessellation by [67])

Figure 2: Tessellation and visits per location
Basemap: ©OpenStreetMap contributors ©CartoDB

rather balanced with a larger number of users and trips and a lower contribution per user. BERLIN has by far the most users and trips distributed on the smallest geospatial area.

5.2.2 Evaluation Runs

To evaluate the similarity of mobility measures with and without privacy guarantees we use different values of the maximal user contribution M and the privacy budget ε. The choices of M are based on the distribution of the number of trips M_u per user u for a given data set: quartiles, the 10th percentile, the 90th percentile, the maximum and the minimum (= single trip per user), resulting in seven variants if no values overlap (e.g., the 10th percentile, second and third quartile have a value of 2 for MADRID). For a given M, a random sample of $\min(M, M_u)$ records is drawn per u.

For comparison, we also include runs that only apply sampling and do not add any noise, which we refer to as *withoutDp*. The proposed differentially private mobility report needs to split the given privacy budget ε between all analyses to grant overall ε-differential privacy. To reduce complexity, we first evaluate each error measure individually, thus the budget is not split, instead the same amount of budget is applied to every single analysis which would sum up for an entire report. In Section 5.5 we elaborate on the implications of privacy budget splitting for the entire report. For evaluation purposes, all combinations of M and ε are run 10 times. In order to present possible deviations, we add error bars to show the variance.

5.2.3 Error measures

We use the following measures to quantify the resemblance between mobility data reports without and with differential privacy guarantees. For any mobility measure x the differentially private counterpart is denoted as x'. The higher the resemblance the lower the error which suggests a better utility of the mobility data report.

TripCountError. We use the relative error to quantify the deviation of total trip counts count_{tr}:

$$\text{TripCountError} := \frac{|\text{count}_{tr} - \text{count}'_{tr}|}{\text{count}_{tr}}.$$

LocationError. The Earth Mover's Distance (EMD) [68] is used to evaluate the deviation of visit counts in each tile. The EMD determines the least amount of work necessary to reshape one distribution to another. Let L be the distribution of visit counts per tile. The EMD between L and L', denoted as LocationError, is defined as:

$$\text{LocationError} := \text{EMD}(L, L') = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}},$$

where n is the number of tiles, d_{ij} is the distance between the i-th and j-th tile and f_{ij} denotes proportion of flows between the two tiles required to transform L to L'. We use the haversine distance between the centroids of tile i and j to determine the distance d_{ij}, allowing us to quantify geospatial shifts, where shifts between neighboring tiles are weighted less than shifts between distant tiles. Thus, we can intuitively interpret the resulting LocationError: A value of, e.g., 100 means that every point in L needs to move 100 meters on average to reproduce L'. Note that we thereby do not evaluate changes in absolute visit counts in tiles, but changes of relative shares.

Table 3: Error measures for user-level vs. item-level privacy with $\varepsilon = 1$. M equals the maximum user contribution per data set (`GEOLIFE`: $2,153$; `MADRID`: 20; `BERLIN`: 16.)

	`GEOLIFE`		`MADRID`		`BERLIN`	
error measure	user-level	item-level	user-level	item-level	user-level	item-level
TripCountError	0.14	0.00	0.00	0.00	0.00	0.00
LocationError	17,142.48	353.50	202.14	18.87	10.32	1.23
OdFlowError	2.00	1.99	1.96	1.78	1.44	0.60
RadiusOfGyrationError	0.50	0.35	0.00	0.00	0.08	0.08

ODFlowError. We use the symmetric mean absolute percentage error (SMAPE) to measure the divergence between matrices representing origin-destination flows. First, the *OD flows* are normalized by dividing through the sum of *OD flows*, so it is not accounted for changes in the absolute count through sampling but only for changes in the distribution. The normalized matrix is denoted as A. To ensure that the error measure is independent of the size of the tessellation, only the combined support of A and A' is considered. Thus, n is the number of combinations where at least one matrix cell of A or A' is not 0. Otherwise, the error could always be reduced by increasing the tessellation and thereby the percentage of 0 valued cells. The absolute percentage is determined and averaged to a single measure defined as

$$\text{ODFlowError} := \frac{2}{n} \sum_{a_c + a'_c > 0} \frac{|a_c - a'_c|}{(a_c + a'_c)},$$

where a_c denotes the entry of the OD matrix A at index c. Note that ODFlowError has a lower bound of 0 and an upper bound of 2.

RadiusOfGyrationError. The error for the radii of gyration makes use of quartiles and is computed using SMAPE. The distribution of the radii of gyration, consisting of the five-number summary, is denoted by ROG:

$$\text{RadiusOfGyrationError} := \frac{2}{5} \sum_{q=1}^{5} \frac{|\text{ROG}_q - \text{ROG}'_q|}{(\text{ROG}_q + \text{ROG}'_q)},$$

where q is the index of the five-number summary value in ROG.

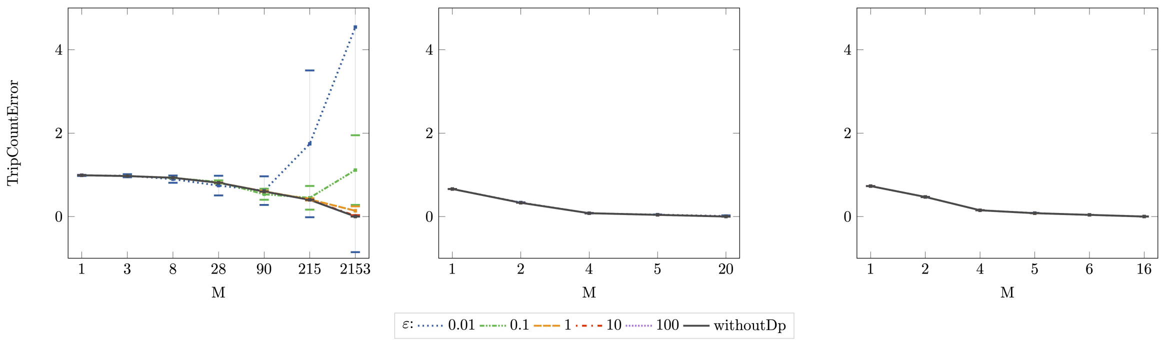
5.3 Item level vs. user level privacy

There is a substantial difference between item-level privacy and user-level privacy for mobility data, as they typically contain multiple items per user. In the following, we denote the maximum number of trips a user contributes to the data set by M, and thus the sensitivity to user-level privacy. Note that in real applications, the provision of M should also be differentially private, unless it is externally set and enforced by sampling techniques [24].

We provide the errors for item-level and user-level privacy for $\varepsilon = 1$ in Table 3. It is hardly surprising that item-level privacy, which provides weaker privacy guarantees, results in less dissimilarity to the original data set than user-level privacy. However, the difference between the privacy levels depends heavily on ε, the data set, and the mobility measure.

The LocationError illustrates well that user-level privacy causes an error several times higher than item-level privacy, in particular for `GEOLIFE` (48 times for `GEOLIFE` vs. 11 times for `MADRID`), which is likely due to the substantially larger sensitivity. On the other hand, adding noise to the total number of trips has almost no impact for either privacy setting. This is to be expected since the measure consists of a single, typically high absolute number that is more robust to noise. Again, for `GEOLIFE`, the high sensitivity strongly affects user privacy guarantees and results in a relative error of 14 % even for this top-level aggregation. On the other hand, there is no relevant difference w.r.t. the RadiusOfGyrationError, as its sensitivity equals 1 and independent of M (the latter holds for most other user analyses, c.f. Table 1). The difference for `GEOLIFE` is likely random, as a standard deviation of $\sigma = 0.20$ for item-level and $\sigma = 0.14$ for user-level indicates.

Finally note that, with the exception of `BERLIN`, even item-level privacy yields an error of more than 100 % for *OD flows*, suggesting that such analyses require $\varepsilon > 1$ to produce useful results. This is due to the fact that the number of origin-destination combinations increases quadratically with the number of tiles, while the corresponding counts are much smaller and thus more sensitive to noise.



(a) GEOLIFE

(b) MADRID

(c) BERLIN

Figure 3: TripCountError for different values of M and ε. The standard deviation of all 10 runs is represented by the error bars.

5.4 Interplay between the upper bound of trips per user M and privacy guarantee ε

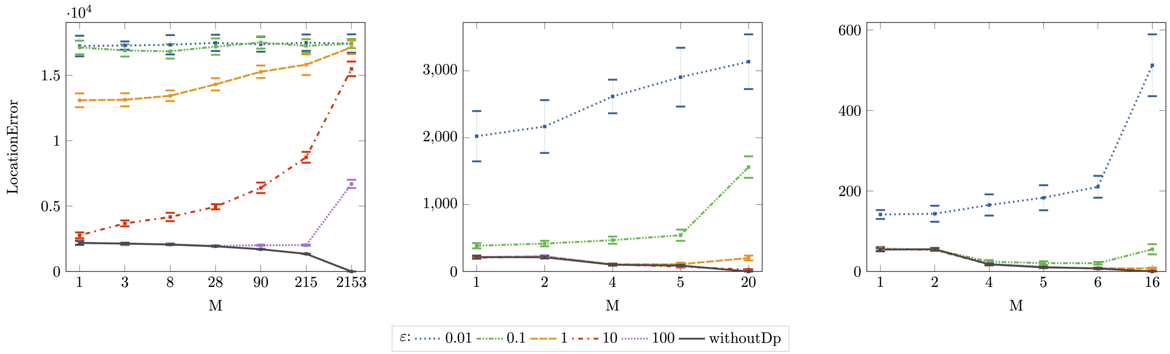
Without the application of privacy measures, sampling of data decreases the similarity to its origin. However, a sampling that bounds the user contributions to a defined maximum M can increase the similarity when user-level differential privacy guarantees are enforced. This is because M defines the sensitivity which in turn defines the amount of added noise. Thus, the question arises which effect outweighs the other: the increase of similarity by keeping more data records per user or the decrease of noise by reducing M. The effect highly depends on the considered mobility measure and data set which we will evaluate in the following. Note that the errors produced purely by sampling without any additional noise, denoted *withoutDp*, function as a lower bound for the differentially private error values.

The evaluation focuses on three aspects: (1) the overall range of error values, (2) the deviation from the baseline *withoutDp*, and (3) the existence of 'tipping points' for error values created through the interplay of M and ε.

As we want to assure that the observed effects are not arbitrary due to the random selection of trips according to M, we first investigate the variance of error measures merely based on different samples. For each value of M, 10 runs are conducted without noise (i.e., *withoutDp*) each based on a new random sample. The standard deviation for each considered error measure and choice of M is low, as all coefficients of variation, i.e., the ratio of the standard deviation to the mean, turn out to be below 0.15. Thus, we can assume that the way how trips were down-sampled for each user has no considerable influence.

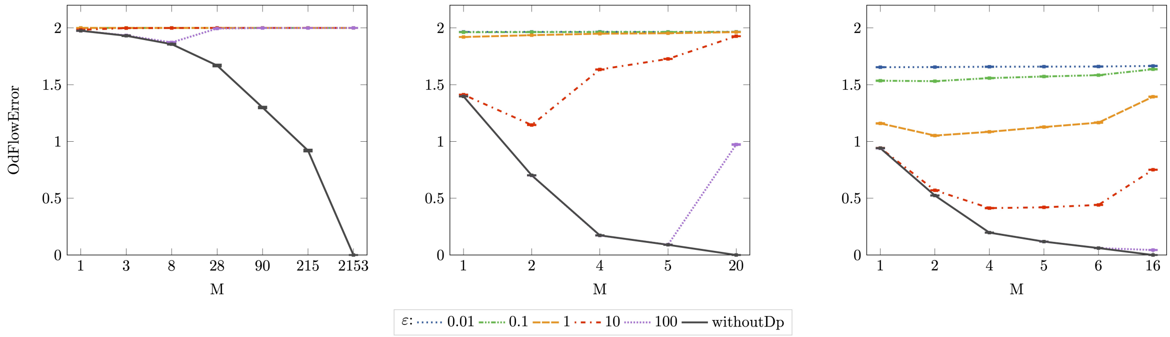
As expected, the effect of additional information dominates that of additional noise for almost all data sets and ε-values for the TripCountError (see Figure 3). Almost all ε-curves resemble the baseline *withoutDp*, i.e., the added noise has no influence on the error. Instead, the error can be attributed to the information loss through sampling. $M = 1$ yields a high relative error between 66 % and 100 % for all three data sets. An absolute count, unlike a relative share, intuitively largely gains similarity to its origin with an increased sample. Accordingly, the curves drop for an increase in M. But even for such a top-level aggregation, a lower sensitivity can outweigh the information loss: for e.g. $\varepsilon = 0.1$, the error for the highly skewed data set GEOLIFE increases from 32 % to 103 % when setting $M = 2,153$ instead of $M = 215$. More fine granular mobility measures where trip counts are disaggregated, e.g., geospatially into tiles, yield smaller values which are in turn more sensitive to adding noise.

In Figure 4 we show the impact of M and ε on the LocationError. Note that the range of the y-axis differs greatly between data sets since the LocationError depends on the underlying extent and split of the tessellation and the uniformity of the geospatial distribution. Recall that the LocationError can be interpreted as the distance every visit needs to be moved on average to create the noisy distribution. The error for *withoutDp* is remarkably low for all three data sets, even for $M = 1$. For MADRID, the error is only 200 meters and for BERLIN $\approx$ 60 meters. For GEOLIFE, the error is 2,290 meters which is rather high compared to the other two data sets. But considering that a sampling of $M = 1$ only contains 1 % (182 trips) of the GEOLIFE data set, one could argue that this error, which corresponds roughly to the diameter of one tile, still lies within a reasonable range. This indicates that there is not a large information gain when increasing M to capture the geospatial distribution, confirming insights



(a) `GEOLIFE` (b) `MADRID` (c) `BERLIN`

Figure 4: LocationError for different M and ε for each data set. The standard deviation of all 10 runs is represented by the error bars. Note, that for clarity the y-axis of the `GEOLIFE` graph uses the scientific notation, i.e., values of the y-axis need to be multiplied by 10^4.



(a) `GEOLIFE` (b) `MADRID` (c) `BERLIN`

Figure 5: OdFlowError for different M and ε for each data set. The standard deviation of all 10 runs is represented by the error bars.

from mobility research stated in Section 4.2 that people mostly visit a few recurrent locations. Overall patterns can therefore already be captured with only a fraction of the actual trips. For all three data sets, the maximum M highly increases the error for most variations of ε, indicating the usefulness of cut off values. E.g., for `BERLIN` the curves corresponding to $\varepsilon \geq 1$ follow the course of *withoutDp*, but for stronger privacy guarantees, such as $\varepsilon = 0.1$, we see a tipping point of the error at $M = 6$. Respectively, for `MADRID` there is a tipping point for $\varepsilon = 1$ and $M = 5$. Again, we see the strongest effect of the sensitivity onto the error for `GEOLIFE`. It is striking that even for a high privacy budget of $\varepsilon = 100$ the error deviates from *withoutDp* at $M = 2,153$. For $\varepsilon = 10$ the error increases for any $M > 1$, while for all $\varepsilon < 10$ the error for $M = 1$ is already a multitude higher than the *withoutDp* baseline, only increasing or staying constant for higher M-values. This suggests that these ε-values are not suitable for meaningful analyses of *visits per location* for `GEOLIFE`.

We present the results for the ODFlowError in Figure 5. Recall that the range for the ODFlowError lies between 0 and 2. `GEOLIFE` clearly does not contain enough data for differentially private analyses of origin-destination flows. For `MADRID` and `BERLIN` much information is already gained with only a small M, as the error for *withoutDp* shows. For $M = 4$ the OdFlowError is down to 17 % for `MADRID` and 19 % for `BERLIN`. Adding noise has a major effect: Even for a large data set like `BERLIN` only a privacy budget of $\varepsilon \geq 10$ falls below 100 %. These results raise the question whether user-level differential privacy guarantees can be given for OD matrices for data sets and tessellations similar in size to those considered here, while still maintaining high similarity and thereby utility of the data sets for further analyses.

5.5 Implications for selection and split of the privacy budget

To create the entire mobility report, an analyst needs to define a privacy budget which must be split between all conducted analyses. As expected, the evaluation showed that the margin of error highly differs for different analyses given the same privacy budget. While the TripCountError remains low even using a small ε, there barely remains any utility for the origin-destination matrix if noise is applied with a similar ε. Thus, to provide a usable mobility report, different considerations needs to be made: (1) To save privacy budget, the selection of analyses should be limited to the ones needed. Therefore, an input parameter is included in the implementation that allows such a selection. (2) Privacy budget should not be split equally between all analyses, thus the implementation provides the option to assign the budget share for each analysis.

6 Conclusion and Lessons Learned

We have compiled typical mobility measures of human movement data to a report and evaluated user-level privacy for selected aggregations on three practice-relevant data sets. Our contribution lays the groundwork for an easily usable open-source tool to create mobility reports with predefined mobility measures and differential privacy guarantees.

We have showed that bounding user contribution has a major impact on error measures in the context of our proposed mobility analyses. The optimal choice of an upper bound M depends on the mobility measure, data set and choice of privacy budget. Bounding user contribution to the 90th percentile of all users' contributions, and thereby down-sampling all 'power users', has the strongest effect on reducing error values for most variations, unless only low guarantees are given by a comparably high ε. In this case there is, as expected, no major impact of the sensitivity, i.e. noise.

Within this work, the same random sample of size M was used to compute all mobility measures. Future implementations should consider different sample sizes for different measures: E.g., absolute counts like the number of trips could be computed with the entire data set, while visits per location could be based on a sample with a small M. The split of the privacy budget between mobility measures should further be optimized as some measures are more resistant to noise. Guidelines for meaningful choices of ε and M would be desirable. It should be noted though that the general question of what privacy budget is suitable is not straightforward and depends on the data set and use case.

Even for supposedly large data sets, counts of single bins shrink to comparatively small numbers as data is disaggregated spatially and temporally. Maintaining a high similarity to the original distribution while preserving privacy of individuals thus becomes a difficult task. Before one optimizes error values, the question should be raised how well a given data set is suited for user-level differential privacy. A data set might not include enough users for meaningful, fine-granular aggregations like origin-destination matrices. E.g., more than 10 % of all trips within the `GEOLIFE` data set are produced by a single user and almost 50 % by 9 users. It is likely that these users do not represent 50 % of a population and its movements to which most use cases using urban mobility data refer. Next, one should be mindful about the desired analyses. A data set might be large enough for temporal and spatial aggregations separately, but not for combined analyses like visits per destination and time or OD flows. Further efforts should be dedicated to optimize the report's utility, e.g., eliminating queries that are not useful for a certain data set. Moreover, by smart privacy budget splitting the utility can be enhanced as the order of magnitude for counts of different analyses vary, thus to balance the margin of errors of all analyses the privacy budget should be allocated accordingly. Additionally, guidance on the size of a tessellation and the number of time windows should be provided.

Our evaluations serve as an orientation for the expected utility for a given privacy budget and vice versa for different mobility measures. Further research should work towards guidelines that are easy to understand and apply in everyday practice. Moreover, the compiled mobility measures should be assessed with practitioners to verify their usefulness. Further measures could be included, e.g., additional variables such as the traffic mode or user demographics. To increase the usability of a mobility report, margins of error resulting from adding noise is thus included in the implementation of the report.

We want to raise awareness to an additional issue that comes with user-level differential privacy: To be able to bound user contribution, a user identifier for each record is necessary. This contradicts the recommendation of data minimization to only store needed information and deserves further discussions in future research.

Acknowledgments

This work is part of the FreeMove project. We'd hereby thank the other project members for their valuable input.

References

[1] Y. Zhou, B. P. L. Lau, C. Yuen, B. Tunçer, and E. Wilhelm, "Understanding Urban Human Mobility through Crowdsensed Data," *IEEE Communications Magazine*, vol. 56, no. 11, pp. 52–59, 2018.

[2] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-Scale Mobile Traffic Analysis: A Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 124–161, Firstquarter 2016.

[3] F. Creutzig, "From smart city to digital urban commons: Institutional considerations for governing shared mobility data," *Environmental Research: Infrastructure and Sustainability*, vol. 1, no. 2, p. 025004, 2021.

[4] R. Ravish and S. R. Swamy, "Intelligent Traffic Management: A Review of Challenges, Solutions, and Future Perspectives," *Transport and Telecommunication Journal*, vol. 22, no. 2, pp. 163–182, 2021.

[5] S. Djahel, R. Doolan, G.-M. Muntean, and J. Murphy, "A Communications-Oriented Perspective on Traffic Management Systems for Smart Cities: Challenges and Innovative Approaches," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 125–151, 2015.

[6] M. J. Nieuwenhuijsen and H. Khreis, "Car free cities: Pathway to healthy urban living," *Environment International*, vol. 94, pp. 251–262, 2016.

[7] B. Faivre d'Arcier, "Measuring the performance of urban public transport in relation to public policy objectives," *Research in Transportation Economics*, vol. 48, pp. 67–76, 2014.

[8] S. Gao, J. Rao, Y. Kang, Y. Liang, and J. Kruse, "Mapping county-level mobility pattern changes in the United States in response to COVID-19," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 16–26, 2020.

[9] A. Aktay, S. Bavadekar, G. Cossoul, J. Davis, D. Desfontaines, A. Fabrikant, E. Gabrilovich, K. Gadepalli, B. Gipson, M. Guevara, C. Kamath, M. Kansal, A. Lange, C. Mandayam, A. Oplinger, C. Pluntke, T. Roessler, A. Schlosberg, T. Shekel, S. Vispute, M. Vu, G. Wellenius, B. Williams, and R. J. Wilson, "Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.1)," *arXiv:2004.04145*, 2020.

[10] J. Pesavento, A. Chen, R. Yu, J.-S. Kim, H. Kavak, T. Anderson, and A. Züfle, "Data-driven mobility models for COVID-19 simulation," in *ARIC '20: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*. ACM, 2020, pp. 29–38.

[11] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, no. 1, p. 1376, 2013.

[12] E. Hargittai, E. M. Redmiles, J. Vitak, and M. Zimmer, "Americans' willingness to adopt a covid-19 tracking app," *First Monday*, vol. 25, no. 11, p. online, 2020.

[13] S. Altmann, L. Milsom, H. Zillessen, R. Blasone, F. Gerdon, R. Bach, F. Kreuter, D. Nosenzo, S. Toussaert, and J. Abeler, "Acceptability of App-Based Contact Tracing for COVID-19: Cross-Country Survey Study," *JMIR mHealth and uHealth*, vol. 8, no. 8, p. e19857, 2020.

[14] O. M. Foundation, "Practical Guide for Cities," Aug. 2020. [Online]. Available: https://github.com/openmobilityfoundation/governance/blob/main/documents/OMF-MDS-Privacy-Guide-for-Cities.pdf

[15] D. Ziemke, I. Kaddoura, and K. Nagel, "The MATSim Open Berlin Scenario: A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data," in *ABM-TRANS '19: Proceedings of the 8th International Workshop on Agent-based Mobility, Traffic and Transportation Models, Methodologies and Applications*, vol. 151. Elsevier, 2019, pp. 870–877.

[16] M. Zakhem and J. Smith-Colin, “Micromobility implementation challenges and opportunities: Analysis of e-scooter parking and high-use corridors,” *Transportation Research Part D: Transport and Environment*, vol. 101, p. 103082, 2021.

[17] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, “Trajectory recovery from ash: User privacy is NOT preserved in aggregated mobility data,” in *WWW ’17: Proceedings of the 26th International Conference on World Wide Web*. ACM, 2017, pp. 1241–1250.

[18] G. Andrienko, N. Andrienko, and G. Fuchs, “Understanding movement data quality,” *Journal of Location Based Services*, vol. 10, no. 1, pp. 31–46, Jan. 2016.

[19] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” in *TCC ’06: Proceedings of the 3rd Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[20] S. Nuñez von Voigt, M. Pauli, J. Reichert, and F. Tschorsch, “Every Query Counts: Analyzing the Privacy Loss of Exploratory Data Analyses,” in *DPM ’20: Proceedings of 15th International Workshop on Data Privacy Management*, vol. 12484. Springer, 2020, pp. 258–266.

[21] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, “Understanding mobility based on GPS data,” in *UbiComp ’08: Proceedings of the 10th International Conference on Ubiquitous Computing*, vol. 344. ACM, 2008, pp. 312–321.

[22] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, “Mining interesting locations and travel sequences from GPS trajectories,” in *WWW ’09: Proceedings of the 18th International Conference on World Wide Web*. ACM, 2009, pp. 791–800.

[23] J. N. Gonzalez, J. Perez-Doval, J. Gomez, and J. M. Vassallo, “What impact do private vehicle restrictions in urban areas have on car ownership? Empirical evidence from the city of Madrid,” *Cities*, vol. 116, p. 103301, 2021.

[24] K. Amin, A. Kulesza, A. Munoz, and S. Vassilvtiskii, “Bounding User Contributions: A Bias-Variance Trade-off in Differential Privacy,” in *ICML ’19: Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019, pp. 263–271.

[25] A. Epasto, M. Mahdian, J. Mao, V. Mirrokni, and L. Ren, “Smoothly Bounding User: Contributions in Differential Privacy,” in *NeurIPS ’20: Proceedings of 34th Annual Conference on Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 13 999–14 010.

[26] Y. Liu, A. T. Suresh, F. X. Yu, S. Kumar, and M. Riley, “Learning discrete distributions: User vs item-level privacy,” in *NeurIPS ’20: Proceedings of 34th Annual Conference on Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 20 965–20 976.

[27] R. J. Wilson, C. Y. Zhang, W. Lam, D. Desfontaines, D. Simmons-Marengo, and B. Gipson, “Differentially Private SQL with Bounded User Contribution,” *PET ’20: Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 2, pp. 230–250, 2020.

[28] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, “Privacy at Scale: Local Differential Privacy in Practice,” in *SIGMOD ’18: Proceedings of the 2018 International Conference on Management of Data*. ACM, 2018, pp. 1655–1658.

[29] A. Hopkins and S. Booth, “Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development,” in *AIES ’21: Proceedings of the 2021 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*. ACM, 2021, pp. 134–145.

[30] J. T. Behrens, “Principles and Procedures of Exploratory Data Analysis,” *Psychological Methods*, vol. 2, no. 2, p. 30, 1997.

[31] N. Andrienko and G. Andrienko, “Spatial Generalization and Aggregation of Massive Movement Data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 2, pp. 205–219, 2011.

[32] J. W. Tukey, *Exploratory Data Analysis*, ser. Addison-Wesley series in behavioral science : quantitative methods. Pearson, 1977, vol. 2.

[33] A. F. Zuur, E. N. Ieno, and C. S. Elphick, “A protocol for data exploration to avoid common statistical problems,” *Methods in Ecology and Evolution*, vol. 1, no. 1, pp. 3–14, 2010.

[34] S. Brugman, “Pandas-profiling: Exploratory data analysis for python,” 2019. [Online]. Available: https://github.com/pandas-profiling/pandas-profiling

[35] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual Analytics of Movement.* Springer-Verlag, 2013.

[36] A. Graser, “An exploratory data analysis protocol for identifying problems in continuous movement data,” *Journal of Location Based Services*, vol. 15, no. 2, pp. 89–117, Apr. 2021.

[37] L. Pappalardo, F. Simini, G. Barlacchi, and R. Pellungrini, “Scikit-mobility: A Python library for the analysis, generation and risk assessment of mobility data,” *arXiv:1907.07062*, 2021.

[38] F. D. McSherry, “Privacy integrated queries: An extensible platform for privacy-preserving data analysis,” in *SIGMOD ’09: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data.* ACM, 2009, pp. 19–30.

[39] N. Johnson, J. P. Near, and D. Song, “Towards practical differential privacy for SQL queries,” *Proceedings of the VLDB Endowment*, vol. 11, no. 5, pp. 526–539, 2018.

[40] C. Dwork and A. Roth, “The Algorithmic Foundations of Differential Privacy,” *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[41] R. Chen, B. C. Fung, B. C. Desai, and N. M. Sossou, “Differentially private transit data publication: A case study on the montreal transportation system,” in *KDD ’12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2012, pp. 213–221.

[42] D. J. Mir, S. Isaacman, R. Cáceres, M. Martonosi, and R. N. Wright, “DP-WHERE: Differentially private modeling of human mobility,” in *Proceedings of the 2013 IEEE International Conference on Big Data.* IEEE Computer Society, 2013, pp. 580–588.

[43] L. Fan, L. Xiong, and V. Sunderam, “Differentially Private Multi-dimensional Time Series Release for Traffic Monitoring,” in *DBSec ’13: Proceedings of the Data and Applications Security and Privacy XXVII.* Springer, 2013, pp. 33–48.

[44] M. E. Gursoy, L. Liu, S. Truex, L. Yu, and W. Wei, “Utility-Aware Synthesis of Differentially Private and Attack-Resilient Location Traces,” in *CCS ’18: Proceedings of the 25th ACM Conference on Computer and Communications Security.* ACM, 2018, pp. 196–211.

[45] V. Bindschaedler and R. Shokri, “Synthesizing Plausible Privacy-Preserving Location Traces,” in *2016 IEEE Symposium on Security and Privacy (SP)*, May 2016, pp. 546–563.

[46] S. Lestyán, G. Ács, and G. Biczók, “In Search of Lost Utility: Private Location Data,” *arXiv:2008.01665 [cs]*, Mar. 2022.

[47] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, “Human mobility: Models and applications,” *Physics Reports*, vol. 734, pp. 1–74, 2018.

[48] L. Pappalardo and F. Simini, “Data-driven generation of spatio-temporal routines in human mobility,” *Data Mining and Knowledge Discovery*, vol. 32, no. 3, pp. 787–829, 2018.

[49] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, “Deep Learning for Human Mobility: A Survey on Data and Models,” *arXiv:2012.02825*, 2020.

[50] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility: User movement in location-based social networks,” in *KDD ’11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2011, pp. 1082–1090.

[51] L. Rossi, J. Walker, and M. Musolesi, “Spatio-temporal techniques for user identification by means of GPS mobility data,” *EPJ Data Science*, vol. 4, no. 1, p. 11, 2015.

[52] S. Bennati and A. Kovacevic, “Privacy metrics for trajectory data based on k-anonymity, l-diversity and t-closeness,” *arXiv:2011.09218*, 2020.

[53] N. Andrienko, G. Andrienko, E. Camossi, C. Claramunt, J. M. Cordero Garcia, G. Fuchs, M. Hadzagic, A.-L. Jousselme, C. Ray, D. Scarlatti, and G. Vouros, “Visual exploration of movement and event data with interactive time masks,” *Visual Informatics*, vol. 1, no. 1, pp. 25–39, Mar. 2017.

[54] E. Toch, B. Lerner, E. Ben Zion, and I. Ben-Gal, “Analyzing large-scale human mobility data: A survey of machine learning methods and applications,” *Knowledge and Information Systems*, vol. 58, 2019.

[55] R. Schlich and K. W. Axhausen, “Habitual travel behaviour: Evidence from a six-week travel diary,” *Transportation*, vol. 30, no. 1, pp. 13–36, 2003.

[56] D. Brockmann, L. Hufnagel, and T. Geisel, “The scaling laws of human travel,” *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[57] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, “Limits of Predictability in Human Mobility,” *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[58] M. C. Gonzalez, C. Hidalgo, and A.-L. Barabasi, “Understanding Individual Human Mobility Patterns,” *Nature*, vol. 453, pp. 779–82, 2008.

[59] T. M. T. Do and D. Gatica-Perez, “The Places of Our Lives: Visiting Patterns and Automatic Labeling from Longitudinal Smartphone Data,” *IEEE Transactions on Mobile Computing*, vol. 13, no. 3, pp. 638–648, 2014.

[60] F. McSherry and K. Talwar, “Mechanism Design via Differential Privacy,” in *FOCS ’07: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*. IEEE Computer Society, 2007, pp. 94–103.

[61] Wes McKinney, “Data Structures for Statistical Computing in Python,” in *SciPy ’10: Proceedings of the 9th Python in Science Conference*, 2010, pp. 56–61.

[62] C. (http://www.crtm.es), “La Encuesta Domiciliaria de Movilidad de la Comunidad de Madrid (EDM2018). data set,” 2018. [Online]. Available: https://crtm.maps.arcgis.com/apps/MinimalGallery/index.html?appid=a60bb2f0142b440eadee1a69a11693fc

[63] Consorcio Regional de Transportes de Madrid, “Documento Síntesis: Encuesta domiciliaria de movilidad de la Comunidad de Madrid 2018,” Nov. 2019. [Online]. Available: https://www.crtm.es/media/712934/edm18_sintesis.pdf

[64] C. (http://www.crtm.es), “Zonificación de transporte ZT1259 de la EDM2018. data set,” 2018. [Online]. Available: https://crtm.maps.arcgis.com/home/item.html?id=97f83bab03664d4e9853acf0e431d893

[65] Y. Zheng, X. Xie, and W.-Y. Ma, “GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory,” *IEEE Data Engineering Bulletin*, vol. 33, pp. 32–39, 2010.

[66] M. Heinrichs, D. Krajzewicz, R. Cyganski, and A. von Schmidt, “Introduction of car sharing into existing car fleets in microscopic travel demand modelling,” *Personal and Ubiquitous Computing*, pp. 1–11, 2017.

[67] G. Berlin, “Verkehrszellen/Teilverkehrszellen in Berlin,” Apr. 2014. [Online]. Available: https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=showMap&mapId=vkz@senstadt

[68] Y. Rubner, C. Tomasi, and L. J. Guibas, “The Earth Mover’s Distance as a Metric for Image Retrieval,” *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.