They See Me Scooting — A Long-Term Real-World Data Analysis of Shared Micro-Mobility Services and their Privacy Leakage

Karina Elzer
Technical University Denmark
DTU Lyngby, Denmark
RPTU Kaiserslautern, Germany
kaelz@dtu.dk

Eric Jedermann
Dept. of Computer Science
RPTU Kaiserslautern
Kaiserslautern, Germany
jedermann@cs.uni-kl.de

Stefanie Roos
Dept. of Computer Science
RPTU Kaiserslautern
Kaiserslautern, Germany
stefanie.roos@cs.rptu.de

Jens Schmitt
Dept. of Computer Science
RPTU Kaiserslautern
Kaiserslautern, Germany
jschmitt@cs.uni-kl.de

***Abstract*—In many places, a surge of micro-mobility sharing systems, as for instance e-scooters, can be observed. Shared micro-mobility is a cost-efficient and flexible alternative to owning vehicles and, furthermore, leads to reduced traffic and air pollution. However, sharing information about vehicles impacts the privacy of the individuals using such vehicles as changes in vehicle state are linked to an individual's mobility pattern. Malicious exploitation of knowledge on mobility patterns of individuals may assist in criminal activities such as stalking or burglary. Thus, it is very important that micro-mobility sharing platforms do not leak sensitive data about the mobility patterns of their users, resulting in a tradeoff between sharing and privacy.**

To characterize the privacy leakage in one specific instance of shared micro-mobility, we conducted a large-scale, long-term data collection from scooters run by the e-scooter company Tier in the European university town Kaiserslautern. Indeed, the data reveals several privacy issues: For instance, we were able to reconstruct work and school schedules of various individuals. Furthermore, we could infer interests and hobbies by visits to, e.g., sports facilities. Our initial discovery of such leakages was aided by the fact that the specific e-scooter company does not comply with existing privacy standards, in particular the use of dynamic IDs. Yet, an a-posteriori analysis of our data shows that even with dynamic IDs, we are able to re-construct 80% of the trips, which still constitutes a substantial privacy leakage.

***Index Terms*—privacy, data collection, e-scooter**

1. Introduction

In recent years, electric kick scooters (e-scooters) have become a popular means of transport. Specifically, the combination of micro-mobility, using a small vehicle, and shared mobility, i.e., renting of a vehicle for a short period of time, is an attractive option. E-scooters in particular are convenient because they are dockerless and small, which allows their placement all over the city. They serve as an alternative to walking if public transport is inconvenient either because of the timetables or distance to the destination.

Many shared e-scooter platforms exist: Bird[1] and Lime[2] have introduced shared e-scooters to cities already in 2017[3]. Since then, multiple other companies followed, such as Tier[4] (becoming dott[5]), Voi[6] and Bolt[7].

Sharing e-scooters also means that the rental companies need to find a way to share the necessary data about available scooters with users and third parties, for instance city administrations. As with any sharing, there is a potential for privacy leakage beyond what users may expect and desire. Consequently, it becomes questionable whether they are consenting to the use of their data and are willing to accept the potential leakage.

Such concerns are mirrored by existing standards. The General Bikeshare Feed Specification (GBFS)[8] is a commonly used standard for sharing availability information of vehicles. It is meant to be freely accessible to everyone. Here, dynamic IDs, i.e., frequent changes of vehicle identifiers, have been proposed as a way to prevent vehicle data from being used to extract user mobility patterns.

We investigate the privacy concern in micro-mobility sharing platforms empirically by collecting data from the scooter sharing system Tier in the European university town Kaiserslautern. The real-world data of available scooters is collected over an extended period of half a year through readily available API access, which provides data about available scooters in the region. It is found that the API does not follow the GBFS standard, in that it includes more data features and does not implement dynamic IDs. We processed the data by extracting relevant features and inferring trips with a focus on the subsequent privacy analysis.

Other works have collected data from micro-mobility scooter sharing platforms as well. However, for several reasons these data collections could not be utilized in our study. Some of the data collections were not conducted for long enough ($\leq$ 3 month) [1]–[4], others relied on too long scraping intervals (5/15min) [5], [6], are no longer available [7], [8] or are aggregated/rounded [7]–[9]. Since we needed a long-term raw data collection with short request intervals, it was necessary to collect our own data.

1. www.bird.co
2. www.li.me
3. https://web.archive.org/web/20241008124416/https://www.theverge.com/2018/9/20/17878676/electric-scooter-bird-lime-uber-lyft
4. www.tier.app
5. ridedott.com
6. www.voi.com
7. bolt.eu
8. https://github.com/MobilityData/gbfs

With our data collection, we are able to identify certain user profiles through clustering trips according to locations and points in time. In particular, we consider privacy-sensitive points of interest (PoIs), such as schools or offices of a particular employer, which allow us to find user groups with specific interests, habits, or living situations. Equipped with these profiles, a clustering for recurring trip patterns and therefore specific users is performed. The results reveal personal information about targeted users, such as their daily routines, their work place and working hours. We validated our approach by obtaining the ground truth for a set of volunteers who consented to having their trips recorded. While the degree of leakage varies between users and is highly related to their frequency of scooter use as well as the locations they visit, our results demonstrate that shared micro-mobility data can indeed be abused for large-scale and efficient profiling as well as stalking of individual users.

Ackermann et al. [10] discuss privacy implications of transport data in various contexts, including shared-mobility data, and suggest practical field studies, however, without actually conducting any such studies. Our paper does exactly that step of going from discussion to practical evaluation for the case of the selected scooter rental service's mobility data and using the GBFS standard. In fact, we find clear evidence for the hypothesis that the chosen anonymization methods and potential additions are not sufficient, using applied data analysis. To the best of our knowledge, Baltra et al. [2] wrote the only other paper with a focus on security and privacy in the rental e-scooter domain. However, they have a very limited dataset with only 10 days and only give one example of identifying people arriving and leaving from a specific PoI, with no recurring pattern analysis. They also do not have confirmation whether the trips were done by the same users, while our work validates it for a subset of known user trips.

On top of that, we check if the dynamic IDs recommended by GBFS could have protected the sharing system against these privacy issues. Dynamic IDs exist in two variants, either the ID of a scooter is randomly changed after each trip or it is changed randomly after a certain duration, which is supposed to make inferring trips difficult. In this paper, we transform our original data set into one that does not use IDs at all, to show that both approaches are still insufficient. We find that a relatively simple heuristic, mainly based on locations and battery levels of scooters, effectively reconstructs the majority of trips. Thus, dynamic IDs are not sufficient to protect users of our specific scooter sharing system and, thus, are most likely an insufficient protection mechanism in the context of shared-mobility systems. A potential protection mechanism is to remove unnecessary identifiers and provide more coarse-grained information on potentially identifying features such as scooter battery level. With such modifications to the data, our concrete inference algorithm is no longer able to reliable identify users in many cases but there are no guarantees that the protection is sufficient.

The proposal of dynamic IDs in the GBFS standard as well as other works suggests that dynamic IDs are expected to notably increase security and privacy [1]. As discussed, our paper exposes this misconception. Baltra et al. [2] is to our knowledge the only other paper mentioning the suspicion that dynamic IDs might not be enough, however, they do not provide any evidence for this.

In summary, we make the following contributions in this paper:

- We collected a long-term, real-world data set from a shared e-scooter provider in a European university town (see Section 4).
- Through in-depth data analysis we demonstrate a severe privacy leakage in that data set (see Section 5).
- A standard solution, dynamic IDs, is shown to be largely ineffective when applied to our data set (see Section 6).

Besides these contributions we discuss related work and the threat model in the following two sections. At the end of the paper, we provide a comprehensive discussion in Section 7: first, potential defenses are examined in Section 7.1; then, we clearly state limitations of our work with regard to the collected data set and its analysis in Section 7.2; to address one particular limitation, we provide initial evidence that our findings generalize to other e-scooter vendors in Section 7.3; finally, ethical considerations are presented in Section 7.4. Section 8 concludes the paper.

2. Related Work

We start by discussing approaches to publishing and anonymizing (shared-)mobility data in general and then detail concrete studies. Last, we quickly summarize existing data sets for e-scooters.

In general, mobility data consists mainly of location and time information. These can be points or traces. It has been shown that spatiotemporal data in combination with identifiers indicating users, can be used to identify the respective individuals: A paper from 2013 [11] uses unicity tests to estimate that only four spatiotemporal points of continuous human mobility traces are sufficient to identify 95% of the individuals. Freudiger et al. [12] analyse location-based services, such as GPS, WiFi and GSM data, connected to unique identifiers. They found that it was possible to infer various points of interest (PoI), including the home and work locations of individuals. Reidentification of users on a large scale with data from call and transportation data sets was demonstrated by Kondor et al. [13] in 2020. Using this data, they uniquely identified 95% of users with only 10 sample points. While these papers show that mobility data in the form of spatiotemporal data can be used to identify users and PoIs associated with the user, they focus on different types of data than in our case: large data sets with multiple data sources and continuous traces. Hence, there is no need to infer trips. Furthermore, the data already contains identifiers that indicate which traces belong to the same user, but not which specific user it is. In contrast, shared-mobility data contains more limited information, such as the start and end point of a trace, which must first be inferred and connected to belong to the same trace. No user identification or information is contained, so trips must first be grouped as belonging to the same user in order to find out further information about this user.

Feature	Our Trips	[3]	[6]	[8]	[4]	[5]	[7]	[9]	[2]
Vendor	Tier	Jump / Bird	Bolt+Fiqsy	Bird+Lime	-	-	-	-	Spin+Bolt+Lift+Bird
City	Kaiserslautern	Washington D.C.	Riga	Louisville	Indianapolis	Berlin	Louisville	Austin / Minneapolis	Los Angeles
Collection Period [month]	6	2.25 / 2.25	7	7	3	9	12	4	1/3
Trips Total	116728	69776 / 13226	750869	79532	-	577812	435,413	661 / 225	-
Trips per Day	637.86	1044 / 226	-	391.78	4830	-	-	-	-
Avg. Distance [m]	1670.28	2138 / 953.3	1364	2140	1800	-	-	0.9 / 1.3 miles	-
Avg. Duration [min]	10.22	17.1 / 15.6	16.82	15.59	13.86	-	-	12 / 19	-
Avg. Speed [km/h]	10.50	-	5.45	9.13	8.78	-	-	5 / 6 [miles/h]	-
Collection Intervals [min]	3	1	5-15	15	-	5	-	-	1-5
Analysis Focus	Privacy	Trip Purpose	Usage	Usage	Usage	Usage	Usage	Usage	Privacy

TABLE 1: Comparison of Related E-Scooter Papers with Data Sets.

Studies of public mobility data demonstrated potential privacy leakage. Two papers [14], [15] on a ride-hailing service show that users and drivers are at risk of having sensitive data leaked. Solutions include rate limits, position concealment through distance, use of synthetic data, removal of identifiers to avoid linkability, privacy-enhancing technologies and cryptographic tools. The difference with e-scooter rental is the addition of the driver and the consequent interaction between driver, user, and company, as well as third-parties, such as cities. Culnane et al. [16] showed that a public transport data set with seemingly anonymized data by changing the card IDs is not sufficient. While the data only contained locations when the user activated the cards, similar to e-scooters, with start and end points of trips, each activation could be connected to a card ID and therefore to the respective user. Re-identification was possible if only two start times and locations of a person were known, or knowledge of similar times or card type in case of co-travelers. While their result is in line with our result that dymanic IDs are insufficient, their data is completely different to GBFS data, so it is not directly applicable in our case.

Different mobility data therefore require different anonymisation techniques as discussed by Ackermann et al. [10]. The authors discussed shared-mobility (Mobility Data Specification (MDS)[9], GBFS), public transport (TRIAS-API), bicycle, walking and parking data. Because mobility data is so diverse, single datatype anonymization techniques are not sufficient to ensure the privacy of the data. As future work, [10] recommends to conduct field studies to address potential anonymization techniques, which we do with shared-mobility e-scooter data.

Xu et al. [1] use three algorithms to infer e-scooter trip data using GBFS. If scooter IDs do not change, they manage to reconstruct and reconnect start and end points. This is done by keeping track of disappearing and reappearing scooters. If IDs do change after each trip, the algorithm allows for inferring but not connecting start and end points. Finally, dynamic IDs periodically allows inference of start and end points by checking whether the number of scooters in a smaller area changes. Our paper uses the same algorithm for non-dynamic IDs to infer trips. The paper's algorithms for dynamic IDs only detects start and end points, leaving possible reconnection of start and end points for dynamic IDs for future work, which we address in this paper. While the authors mention privacy as a reason for dynamic IDs as an aside, their focus and reason for inferring trip data is usage analysis. Our paper focuses on the privacy aspect of the trip reconstructions, even under dynamic IDs, proving their statement wrong.

9. https://github.com/openmobilityfoundation/mobility-data-specification

Baltra et al. [2] examine the privacy of shared-mobility standards. While the MDS is not intended to be public, the historical trip data leads to the user's privacy not being protected from the cities and third-parties that have access to the data. Baltra et al. argue that GBFS is enough for cities to analyse usage and distribution. Furthermore, the paper shows with an example that GBFS with non-dynamic IDs is not sufficient. Their solution is geo-indistinguishability. In their example, they reconstructed trips by clustering with K-means to identify visits to a marijuana dispensary. However, they only go as far as identifying different consecutive trips to and from one specific sensitive area by the same user. They do not show how it is possible to refine profiles for specific riders or use auxiliary data to find potential personal habits of a user, including residences or work locations and other habits. We show that our approach can achieve this and works on a large scale, on a town with several thousands of citizens.

Differential privacy [17], which adds noise to data to hide data of individuals, has been suggested in the context of hiding locations [18]. However, as vehicles move on clearly defined roads, it is not possible to simply add noise as that would result in unlikely positions, which can be traced back to the original position. There exist more sophisticated solutions that ensure that vehicles end up on roads. However, they are designed for tracking goods and only require a very low accuracy. For e-scooters, the location information is required to be close to accurate as users have to be able locate a scooter they want to use [19].

In the context of e-scooters, there exist several works about collecting and using a data set of e-scooter trips. The data sources vary from API scraping [3], [5], [6] to Open Data platforms of cities [4], [7]–[9]. The data sets consist of different e-scooter vendors, features, collection time, city, and consequently size. See Table 1 for a comparison of our study with related work dealing with e-scooter rental platforms. In particular, the focus of the analysis shows that the majority of the papers aimed at analyzing the usage. As Table 1 shows, the datasets of these papers have a small collection period or large collection intervals. In addition, since usage analysis does not require precise times and locations, some of the datasets are aggregated [7]–[9]. The only other privacy paper [2] has a dataset of only 10 days, which makes long-term privacy leakage analysis impossible.

3. Threat Model & Privacy Concerns

In this work, we evaluate whether and how easily privacy-sensitive user-related information can be extracted from shared mobility data of e-scooters.

Some user-related private information that may be revealed are: an individual's residence, company and work or school schedules, personal interests through the visits of sport clubs or cultural events, religious beliefs through visits to churches, synagogues or mosques, or health problems that come with regular visits of medical specialists or psychiatrists.

An obvious prerequisite to be under this privacy threat is the frequent, or even regular usage of scooters in one's everyday life. In this way links are created between individual daily routines and trips in the sharing system, which can be observed. Thereby, it is possible to derive information about users' routines.

3.1. Attack Scenario

In the given scenario an adversary is able to gather scooter data and later on analyze it. This information is publicly available, since rental companies provide information about their scooters. So on the technical level the adversary is an active participant, requesting the data from the provider's website. On the conceptual level, the adversary is an eavesdropper, only passively observing the usage of scooters without actively interfering with the system.

It is possible to retrieve privacy-sensitive information about a target in the gathered data in two ways: either by defining a target profile to search for possible users or by starting with a known residential address of a target user. The first approach via defining a target profile can be applied when searching for a member of a specific group, for example an employee of a company, a member of a religious community or a child. Each example can be realized by searching for trips that start or end at the entrance of a factory, a church, or a school. By finding recurring trips from or to such a point of interest (PoI) that start or end in a residential area, the potential home address of a user with the correct profile can be inferred.

From there on the second approach of starting with a known residential address can be applied. Via analyzing the incoming and outgoing trips of this address it is possible to gather additional information about the user(s) associated with this address. It may be possible to extract how many people live at the given address, find out potentially sensitive details about the daily schedule of a scooter user and gain information about someone's hobbies and habits.

3.2. Assumptions

This work makes several assumptions about scooter users and their usage:

First, we assume that scooters are used to drive directly to a desired destination. Since scooter sharing systems are a popular solution to address the last mile problem of public transport [9], they are a convenient possibility to travel and can be parked nearly everywhere. So it should be reasonable to assume that the end of a scooter trip is often the desired destination of the user.

Second, it is assumed that users departing from a location are searching for the closest available scooter and use this one for their next trip. Nevertheless, this leads to departures being more spatially spread in comparison to destinations.

The first two assumptions allow us to identify addresses that are PoIs of scooter users. A third assumption is that users have routines, which can be detected by combining frequent occurrences of feature values such as similar starts, destinations and times of the day. This is based on the assumption that most people have a scheduled day with recurring activities. This allows us to associate multiple trips with presumably one user.

A fourth assumption is that addresses with clear association with a single entity or purpose reveal some information about the users intention to target this location. The third and fourth assumptions allow to draw conclusions on the routine and the intention of a user. For example, if the destination of a daily trip-cluster is an address in a region with only detached houses, this allows us to identify potential residences of scooter users. If the start of this cluster is in an industrial region that is clearly related to a specific company, this likely reveals the workplace of the scooter user.

As previously mentioned, the adversary must have access to the scooter data and no rate limit for requesting.

4. Data Collection & Processing

This section presents the scooter rental's API and how the data was collected. The data reveals various properties of the available scooters, which we call features. These features are used to process and extend the data such that we can use it for our purpose of privacy analysis. Furthermore, a short overview of general feature statistics are given to show what entries the final data set consists of. Finally, an alternative way to gather more detailed data by live-tracking scooters is briefly discussed.

Note that the scooters are dockless, meaning they can be stationed anywhere. This rises the necessity for the vendor to occasionally relocate scooters from remote locations to central transportation hubs.

4.1. API

E-scooter rental companies typically provide an API endpoint that allows the users to see the currently available e-scooters that can be rented. For our rental service Tier, this is the 'vehicle' endpoint. An outdated and incomplete documentation can be found online[10]. In order to query this endpoint, a generic API key is required, which can be found by observing the apps traffic when logged in. Alternatively, an API key is available on a GitHub page[11] not affiliated with the rental service, which allows for API usage without any connection to one's personal account. This results in the following query:

"https://platform.tier-services.io/v2/vehicle?zoneId=KAISERSLAUTERN"

The API endpoint provides only information about scooters, no information about user accounts, trips, or payments can be directly requested. The request returns a list of JSON objects, each representing scooters available at the

10. https://web.archive.org/web/20240314114227/https://api-documentation.tier-services.io/

11. https://sharedmobility.github.io/Tier.html

time of the request using 18 attributes (Listing 1). Notably, these attributes include a static ID, exact coordinates, battery level in percent and a timestamp of the last location update. The number of scooters returned depends on the number of available scooters at that moment in time, scooters that are in use are not returned by this zone query, and, therefore, disappear from the dataset and reappear once they are available again.

Listing 1: Example Entry of an Available Scooter in the APIs JSON Response

```
{ "type": "vehicle",
"id": "1b096d37-d444-49ba-bbaa-adf44dc967ee",
"attributes": {
"state": "ACTIVE",
"lastLocationUpdate": "2023-09-25T20:15:43Z",
"lastStateChange": "2023-09-14T20:55:36Z",
"batteryLevel": 43,
"currentRangeMeters": 16000,
"lat": 49.454534,
"lng": 7.765172,
"maxSpeed": 20,
"zoneId": KAISERLAUTERN,
"code": 143134,
"iotVendor": "okai",
"licencePlate": "911WRX",
"isRentable": true,
"vehicleType": "escooter",
"hasHelmetBox": false,
"hasHelmet": false }}
```

4.2. Collection

We implemented an automated script that allows us to send requests to Tier's API in regular intervals to gather real-world data of available scooters over a long-term period. We used a Raspberry Pi as the server to run the Python script with additional notifications through Discord in case of issues. The data collection was done every 3 minutes over a period from October 1 2023 to March 31 2024, covering exactly half a year. These parameters were selected due to restrictions on time and storage as well as to avoid overloading the servers. The zone was limited to Kaiserslautern, because of storage limitations and our knowledge of the area. The town has an area of roughly $140km^2$ and a population of approximately 100,000.

In total, this resulted in 80980942 entries of available scooter data points being collected (see also Section 4.3), while 1223 unique scooters were observed. An example of what an entry of an available scooter looks like can be found in Listing 1. The raw data is stored in a JSON file for each request, with the total volume of data stored being 38.2 GB. Furthermore, an SQLite database was created for easier querying, and adjustments were made to deal with the daylight saving time, which is relevant during the observation time frame.

4.3. Features

Of the 18 different features for the e-scooter found in the JSON responses, 6 are relevant for our purposes. These are the unique scooter ID, last location update timestamp, latitude and longitude, last state change timestamp, and the battery level. Additionally, each entry includes the time of the API request, which we call collection time. It should be noted that without the ID, the licence plate and vehicle code could be used as an alternative unique identifier.

The type, state, maximal speed, zone ID, IoT vendor, vehicle type, and features related to helmet box are all expected to be static. However, the data set contains 257 entries with a different vendor than the others, as well as unusual coordinates. These entries are therefore removed. The helmet related features are ignored, as they are not relevant for the analysis, but both 'false' (expected value) and 'true' appear as values in the data set for a small number of entries. This is unexpected and shows that the API is not entirely reliable. While the documentation states that a state change happens at the start and end of a trip, this could not be confirmed with test drives. We conducted two drives: one shorter of 5min and a longer one of 15min. Based on the test drive evidence, the features regarding rental and state change were neglected for the event definitions. Figure 1 shows the actually observed process above the timeline, while the process described in the documentation is displayed below the timeline. Our method is based on the observed behavior.

4.4. Processing

To infer trips, the entries are compared to previous entries with the same ID. Consequently, the first entry of each ID is removed, because there is no previous entry. This is based on the fact that only available scooters are returned by the API, while scooters in use are missing. For a non-available scooter this means, the entry before and after it was not available will be captured. While we do not know what happened during a renting, we know where the scooter was before and after it. The combination of entries allows calculation of additional features for the created events: duration, battery change, distance, battery change per distance, and speed. The resulting entries are all events, so that not moving or activating a scooter still is classified as an event. To distinguish further, we specifically refer to trips, roundtrips, loadings and relocations as active events.

For distance, the fastest route from the start to the end point is calculated with OSMnx[12] for bicycles. The bicycle option is chosen since it is expected to be the most similar in terms of speeds to scooters in comparison to walking and driving, the only other two options provided by OSMnx. The duration is calculated based on the observed behavior, as depicted in Fig. 1. The timeline shows that the best possible start time to choose is the previous collection time, so the last time the scooter was observed before it disappeared from the data set, this can lead to a margin of error of up to 3 min. An alternative would be the last location change returned by the API, which could lead to a bigger margin of error, since it will be either the same as the collection time or before. The best possible end time that can be selected is the last location update, which avoids the 3min error for the end of a trip.

Based on observations in related work [3], [5], [6], which analyze available scooter data, disappearing scooters do not necessarily have to be trips. Rather, it can also be an indication for loadings, roundtrips and relocations.

12. https://osmnx.readthedocs.io/en/stable/

Feature	Trips	Roundtrip	Relocation	Loading
Distance [m]	1670.28	39.97	1794.18	206.64
Duration [min]	10.22	15.70	167.95	801.48
Speed [km/h]	10.50	0.41	12.72	0.46
Battery Change [%]	-4.24	-0.50	-1.44	+23.96
Battery Change per Distance [%/km]	-2.45	-19.16	-0.71	+1632.26
# per Day	637.86	359.51	189.14	320.45
# per Hour	26.58	14.98	7.88	13.35

TABLE 2: Average Values of Features for the Active Event Types.

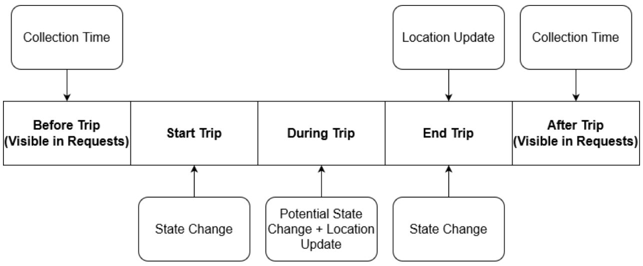


Figure 1: Trip Timeline Based on Observation (above) vs. Documentation (below).

Those events have their own characteristics and need to be distinguished. We take the same definitions, but choose our own cut-off values. Trips, for us, are any events with duration (dur) longer than 4 minutes, which means the scooter must have disappeared from the data set. Otherwise, the scooter was standing or moved while not activated. Furthermore, the duration is not allowed to be above 120 minutes, since our rental company mandates that scooter trips are not allowed to be more than two hours. A negative or zero battery change (bc) can be used to distinguish them from loadings. Compared to roundtrips, trips should have a distance (dist) greater than 500 meters. We chose 500 meters after analyzing the distribution of all distances. Movements below 500 meter did not have a clear pattern and were very rare, which can be explained by the initial unlocking fees of scooters, making very short trips unattractive. Roundtrips have a small distance, since the distance is only an estimation based on the start and end location. Finally, the maximum speed for trips is set to 17km/h, because the scooters can go 20km/h maximal and on average this will not be reached. The battery change per distance (batt_dist) to a maximum of -1%/km, unless the battery has not changed and the trip is at most 1km long. The maximum of -1%/km is chosen here, because we assume during a trip the battery level will change, however depending on certain factors, such as the terrain, it might not be much. It is therefore the upper bound that is used to make sure, every trip is correctly filtered. These boundaries were set to differentiate trips from relocations, since relocations are assumed to be done by car, therefore they are faster and have no or hardly any battery change, as the e-scooter is not used. This results in the following filters, which can be applied in the following order to identify the events:

- Loading: scooter batteries are loaded, leading to positive battery change.
  $(bc > 0)$
- Standing: the scooter was not rented or moved.
  $(bc \leq 0 \wedge dur < 4 \wedge dist = 0)$
- Moved: the scooter was not rented but changed location because it was moved by hand (or GPS inaccuracies).
  $(bc \leq 0 \wedge dur < 4 \wedge \neg(dist = 0))$
- Roundtrip: the scooter was rented (disappeared), but the end of the trip was close to the start of the trip.
  $(bc \leq 0 \wedge dur \geq 4 \wedge dist < 500 \wedge dur \leq 120)$
- Trip: the scooter was rented (disappeared) and based on the features distance, duration, battery change, speed limits (as explained in more detail the previous paragraphs) it is realistic that the results are a trip, were a person actually drove the scooter.
  $(bc \leq 0 \wedge dur \geq 4 \wedge dist \geq 500 \wedge speed \leq 17 \wedge dur \leq 120 \wedge (batt\_dist \leq -1 \vee (bc = 0 \wedge dist \leq 1000)))$
- Relocation: the scooter disappeared and based on the features it could not have been a normal ride, but rather other transportation was used to move the scooter, for example because of maintenance or changing placement to a location with higher usage.
  $(bc \leq 0 \wedge dur \geq 4 \wedge dist \geq 500 \wedge (speed > 17 \vee dur > 120 \vee (batt\_dist > -1 \wedge (bc \neq 0 \vee dist > 1000))))$

Furthermore, certain events were removed as outliers. Specifically, those with coordinates way outside the town, a speed greater than 20 km/h and events with a duration greater than 120min, but less than 500m distance and negative battery change, as these do not fit to loadings, trips or relocations.

4.5. Analysis

Table 3 shows the resulting numbers of all the different event types. Due to the nature of our dataset being snapshots of available scooters, the filtering into different events was necessary. The majority of all events are passively standing and very minor movements, which is expected since most of the times the scooters are standing, waiting to be used, with minor movements being due to them being moved out of the way or being a result of GPS inaccuracies. In total, the data processing leads to 116728 trip events or, in other words, 0.14% of all events are recognized as trips. However, the sum of active events (trip, roundtrip, loading and relocation) are 0.34% of all events. Inside this active group the trips are the majority with $\sim 42\%$.

The averages of the most important features for the active events can be seen in Table 2. Especially roundtrips with a battery change of -19.16%/km stand out. This

Event Type	# Events	Total Events [%]	Active Events [%]
Standing	77801631	96.08	-
Moved	2899253	3.58	-
Trip	116728	0.14	42.33
Roundtrip	65791	0.08	23.86
Loading	58643	0.07	21.26
Relocation	34612	0.04	12.55
Outlier	3040	0.004	-
All	80979719	100	0.34

TABLE 3: Number of Events per Type.

can be explained by the inaccurate distance calculation. Having only start and end points for routing, does not allow better results for roundtrips. The low battery change is most likely due to prematurely canceled trips, were the scooter might not have worked or similar, and therefore did not have a large distance and is counted as roundtrip. The speed value for loading, in comparison with the average roundtrip speed, suggests that loading happens at times by changing the battery on site, making the routing distance for loadings inaccurate as well. Furthermore, the average duration suggests exceptions to loadings on site, most likely due to maintenance in the warehouse.

Specifically for trips, the mean of the different features and a comparison to other papers can be found in Table 1. The results show that the trip definition is sufficient to lead to similar averages to those of other studies, while keeping in mind that e-scooter usage may vary depending on demographic and geographic variables [7], [9].

4.6. Live Tracking

As previously explained, the 'vehicle' endpoint is only supposed to return currently available e-scooters. When querying a zone, such as a city, this is the case. However, when querying a specific scooter ID, the information of the scooter will be returned even if it is in use. Here the endpoint "/v1/ vehicle /SCOOTER_ID" was used. Instead of a list, it returns the same style JSON object (Appendix 1) for the requested scooter, no matter if this scooter is available or not. This makes live tracking of scooters possible. With additional information that links a scooter to a rider, it is therefore also possible to track users. Hilgert et al. also mention this in a paper from 2021 [20]. We were able to confirm that this statement is still true. A simple algorithm can be used to detect new trip starts, and then track the vehicles used until the trip ends. Starts and ends can be detected by observing whether an e-scooter disappears or appears, respectively. Knowing the ID of the disappearing entry, a specific request for this scooter can be made in regular intervals until it reappears in the zone query.

A proof of concept of this tracking algorithm was implemented and tested. The result of a short example trip that was taken at the university is shown in Figure 2. The blue path shows the GPS track that was captured during the ride and the red shows the path tracked by the algorithm. It is clear that the path returned by the tracking algorithm is slightly less accurate than the GPS track. This is because of the time interval between requests of the position, which was changed to every 10 seconds for the live-tracking algorithm as a good compromise between overloading the API and getting a good live track.

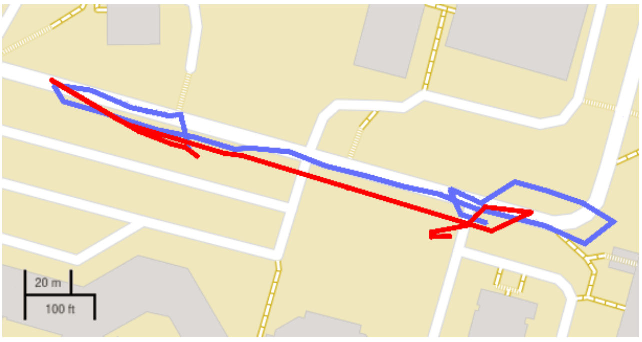


Figure 2: Example Tracks by Live-Tracking Algorithm (red) and GPS (blue).

Nevertheless, the results of the tracking algorithm show the general path of the vehicle, as expected. This confirms the possibility of live tracking.

This method could lead to the possibility of tracking down and following specific individuals only by knowing the scooter ID or license plate or even just the location or time from physically anywhere. It could be used for stalking or also for knowing when no one is home, in real time.

However, this does not just allow live-tracking, but could also be used to gather more detailed trip data, including routes and therefore more accurate feature data. We chose to not collect data this way for multiple reasons. It would require more storage and resources for the additional requests and entries. Depending on the request interval, which would need to be higher for more accurate trip data, it could lead to significantly more request. More requests might lead to issues because of a rate limit mentioned in the documentation. For avoiding such problems and also regarding the ethical concerns, mentioned in Section 7.4, we decided to not perform such an intense observation.

5. Privacy Evaluation

The goal of our privacy evaluation is to analyze if publicly available scooter data holds information about recurring and frequent trips that strongly indicates trips to be done by the same person. Therefore, we extract the trips of each scooter, as described in the previous section. To find trips presumably executed by the same person, we use spatial and temporal clustering. Spatial clustering allows to identify clusters of trip starts and destinations. As already discussed in Section 3.2, rentable e-scooters can be parked nearly everywhere, so users are assumed to pick up the next available scooter, ride it directly to their destination and park the scooter there. Both actions, the pick up and parking of a scooter are observable in the scooter data with respect to their location and time. Recurring trips, from the same start to the same end point, performed at a similar time of day, are the basis for our investigation. In the following, we first describe our approach and then present some notable user profiles that we found. At the end, in Section 5.5, we present a validation of our approach on a data set from volunteers that consented to sharing their trip data.

5.1. Integration of OSM

To integrate map information into the trips, we combined the start and end locations with data from open street map (OSM). OSM provides information about building types and land use. Thereby it is possible to filter areas of residential usage with keywords such as “detached house”, “semidetached house”, “apartments” and “dormitory”. Some other land use categories are commercial, including “bar”, “dentist” and “supermarket”, or public, including “townhall”, “university” and “school”. Using this information, it is now possible to identify trips that are close to potential residences of scooter users. There are residential building types that only reveal limited information, e.g., dormitories, which often hold dozens to hundreds of students. Consequently, there are likely a high number of people commuting from these buildings and we cannot easily determine which person is associated to which trip. But other buildings reveal more reliable information about scooter users: Trips that start or end close to detached or semidetached houses suggest that the driver is one of a few family members. This reduces the number of candidates to perform trips and often allows to connect trips to individuals. Also locations in public areas reveal different levels of information: supermarkets, which are visited by a large variety of people each day do not reveal so much, while destinations close to a school lead to the conclusion that the scooter is used by pupils or teachers.

Thus, integrating OSM enables linking start and end locations of a trip with a trip purpose. This also allows to distinguish between target groups.

5.2. Profile Clustering

In this step of the privacy analysis, the goal is to find potentially interesting scooter users with a certain profile. The profiles that we can find are linked with visiting associated locations, such as employees working at a factory, sport club fans cheering at a stadium, or kids visiting a school. So for finding potentially interesting scooter users, we are searching for trip clusters that start or end at a location that is associated with the profile that we are searching for.

The features utilized for clustering are start and end location, as well as time of day. Each of these features is clustered separately. For spatial clustering we utilize Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [21]. This allows the generation of clusters based on the minimal number of points ($MinPts$) for one cluster and the maximal distance of one point to the next point (ε), without predetermining the total number of clusters.

$MinPts$ is set to 6 for both start- and endpoints. We choose the number 6 since we observed a total period of 6 month and the intention is to be able to find regular trips that happened at least once per month. We choose different values of ε, ε_S and ε_E, for start- and endpoints. To cluster the endpoints we are targeting for smaller, denser clusters, based on the assumption that users drive directly to their destination. So for ε_E we tested a range of 1 to 10 meters and eventually decided to use ε_E = 5 m. For clustering start points, we initially intended to use a larger distance of 10 to 100 meters, since users have to find the closest scooter and thereby select scooters from a larger region. However, we came to the conclusion that using smaller ε_S of 7.5 meter produces better results. For larger distances the clusters often become too large, including many falsely clustered scooters, which would have significantly reduced the value of the clustering results.

To cluster trip times we employ the k-means algorithm [22], which requires a selection of the value for the parameter k. We decided to split the day into k intervals. We compared the clustering results for $k \in \{12, 24, 48, 96\}$ and finally $k := 48$ is selected. This results in an average duration of 20 minutes. To overcome the problem of the clock-transition at midnight from 23 o'clock to 0 o'clock, we apply a simple sine and cosine transformation to map the 24h day on a unit circle and use those two values for clustering. The following are the corresponding formulas:

$$time_{sin} = \sin\left(2\pi \cdot \frac{time\,[s]}{86400}\right)$$
$$time_{cos} = \cos\left(2\pi \cdot \frac{time\,[s]}{86400}\right)$$

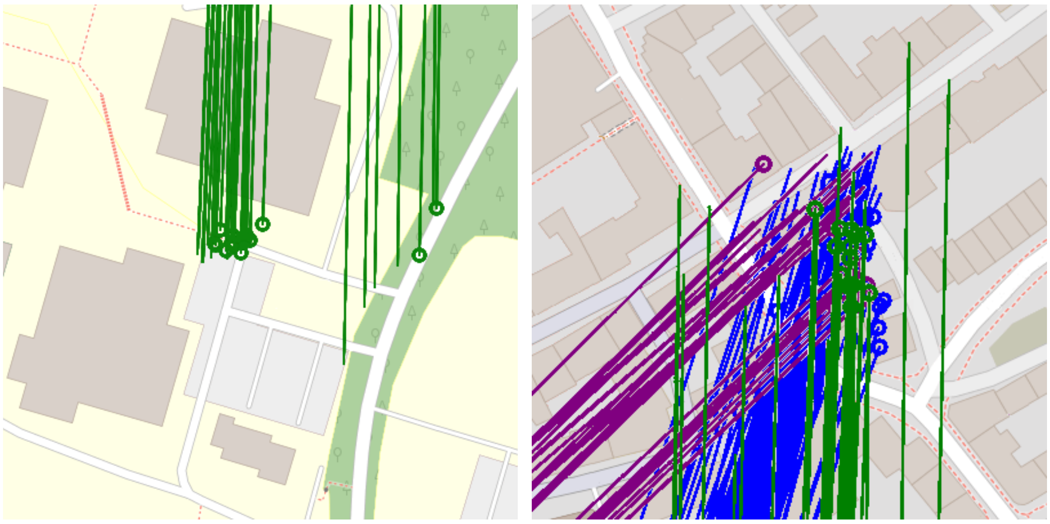
The single trips can now be combined in trip-clusters, where each trip starts in the same start cluster, ends in the same end cluster, and lies in the same time cluster. Using the assumption that people tend to have a regular schedule, it can be assumed that trips that start and end in the same region and are performed regularly around the same time are likely by the same person. Since we start at a location of a certain interest, we can infer that a person regularly visiting this place also fulfills the profile we are searching for.

At the same time, we can use the cluster data at the location of interest to gain some insights about local businesses. For instance, many trips arriving or departing at the same time reveal something about the start and end of working shifts or school time schedules.

5.3. Residential Clustering

After filtering the trip clusters by location, we obtain trips with start and end positions that are likely associated with the same user. Given the clusters, we can use the spatial proximity of other trips' start and end locations to gather more data for that specific user. We start by considering locations in residential areas that do not have a high scooter usage. This allows us two things: first, we can identify potential home addresses of scooter users; second, we can hypothesize that other trips that start/end in the a location cluster close to the same address are likely done by the same person. In this way, we can combine more trips based on spatial proximity to obtain even larger amounts of user-related, potentially privacy-sensitive data.

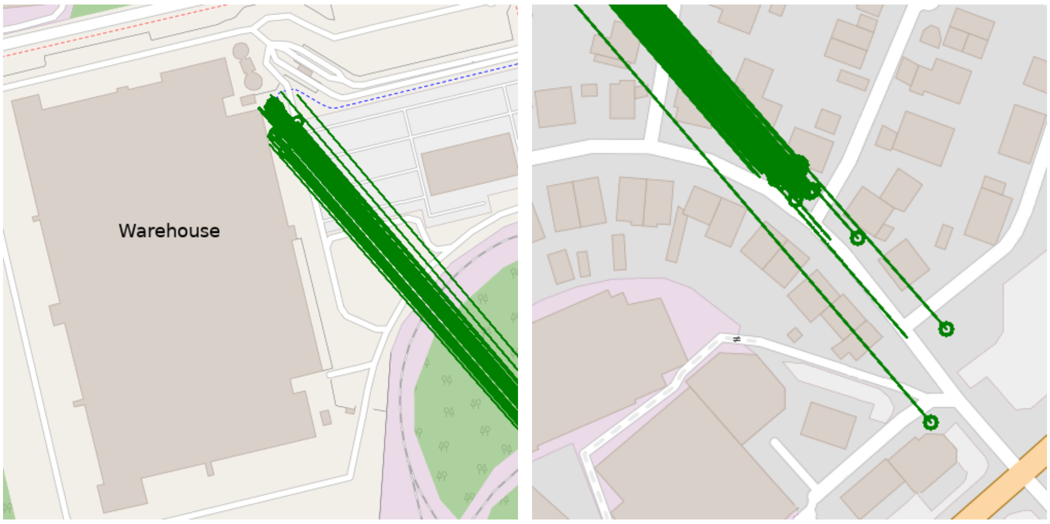
We conducted this step mostly with potential home addresses in residential areas, which turned out to be a more reliable strategy than using crowded locations with higher fluctuations of users and scooters like the town center or the university campus. Thereby, it is possible to gain insights about hobbies, friends, or family members.



(a) Trips from and to school.

(b) Trips from and to residence.

Figure 3: Trips of a multiple people living in the same house: Green: School, Blue: Train Station, Purple: Supermarket. (destinations are indicated by a circle at the end)



(a) Trip ends at employer.

(b) Trip ends at residence.

Figure 4: Commute Trips of a Warehouse Employee.

Date	Weekday	Arrival at school	Arrival at home
10-10-2023	Tue	07:48:45	12:52:48
02-11-2023	Thu	08:01:15	12:58:20
31-10-2023	Tue	07:53:31	
07-11-2023	Tue	08:02:06	
09-11-2023	Thu	08:12:45	
28-11-2023	Tue	07:46:41	
12-12-2023	Tue	07:57:48	13:23:30
23-01-2024	Tue		13:01:11
25-01-2024	Thu	08:15:35	12:49:23
30-01-2024	Tue	08:10:11	13:14:17
01-02-2024	Thu		12:30:23
15-02-2024	Thu	07:54:38	11:30:06
22-02-2024	Thu	07:56:17	
05-03-2024	Tue	07:57:34	
14-03-2024	Thu	08:03:17	12:24:41
19-03-2024	Tue	07:58:15	

TABLE 4: Tuesday & Thursday School-Trips.

5.4. Privacy Revelation

To get a first impression on how much privacy leakage may result from shared scooter data, we filtered regions associated with 4 categories of everyday life and counted the number of residential addresses revealed by trips starting or ending in these regions:

For 7 secondary schools, we found 14 residential addresses, where regular trips towards schools were made in the morning right before 8 a.m. At all three public swimming pools, we found in total 11 residential addresses, despite analyzing the winter term, when the outdoor pools are partly closed. Regarding police and military, we found 7 addresses. Regarding commuters to some large companies (warehouse, a hospital, a research center, one arms factory), we found in total 24 addresses.

In the following, we describe a more in-depth analysis of a school associated user, to emphasize how much information can be gathered. During the 6 months of

Date	Weekday	Trip End Time	Direction
20-12-2023	Wed	07:01:40	way home
20-12-2023	Wed	21:46:44	trip to work
21-12-2023	Thu	06:56:15	way home
21-12-2023	Thu	21:45:41	trip to work
22-12-2023	Fri	07:11:49	way home
22-12-2023	Fri	21:44:52	trip to work
23-12-2023	Sat	07:12:23	way home

TABLE 5: Commute Times of an Warehouse Shift Worker.

Date	Working shift	Shift times	available data
11-12-2023 (Mon) - 16-12-2023 (Sat)	morning shift	6:00 - 14:30	2x to warehouse, 4x way home
20-12-2023 (Wed) - 23-12-2023 (Sat)	night shift	22:00 - 6:30	3x to warehouse, 4x way home
25-12-2023 (Mon)	afternoon shift	14:00 - 22:30	-
3-1-2024 (Wed) - 6-1-2024 (Sat)	morning shift	6:00 - 14:30	2x to warehouse 3x way home
8-1-2024 (Mon) - 10-1-2024 (Wed)	night shift	22:00 - 6:30	1x to warehouse, 2x way home
15-1-2024 (Mon)	afternoon shift	14:00 - 22:30	-
22-1-2024 (Mon)	morning shift	6:00 - 14:30	-
29-1-2024 (Mon)	night shift	22:00 - 6:30	-
8-2-2024 (Thu) - 10-2-2024 (Sat)	afternoon shift	14:00 - 22:30	2x to warehouse, 2x way home
12-2-2024 (Mon) - 17-2-2024 (Sat)	morning shift	6:00 - 14:30	2x to warehouse, 2x way home
20-2-2024 (Tue) - 23-2-2024 (Fri)	night shift	22:00 - 6:30	1x to warehouse, 3x way home
15-1-2024 (Mon)	afternoon shift	14:00 - 22:30	-
5-3-2024 (Tue) - 9-3-2024 (Sat)	morning shift	6:00 - 14:30	3x to warehouse, 4x way home

TABLE 6: Reconstructed Shift System of a large Warehouse Operator.

data collection, we found between a few dozen to many hundreds of scooter trips for each secondary school, with schools being closer to the town center being associated with more trips.

Whenever school attendees (staff or kids) regularly use scooters for their way to school, they reveal some information about their daily schedule. In one such case we found a household with three persons, possibly two adults and one school kid. The trips' start and end locations at their residence are shown in Figure 3b, with the trips to school shown in green. The travel time for one school trip is about 10 to 15 minutes. The scooters are used 28 times for the way to and from school, as the trips in Figure 3a show. The lines in the graphs connect the start and end of a trip, the destinations are indicated by a circle at the end of the line. In Figure 3a there are two trips visualized with a circle that end directly at the street and a handful of trips starting at the street. All other trips start and end in a cluster on the schoolyard. 23 of the total of 28 trips are done on Tuesday and Thursday, this snippet of the data is shown in Table 4. For some days, like December 12 and January 30, a school-typical pattern with rides to school around 8 a.m. and back from school shortly after 1 p.m. is clearly visible. For all other days the arrival time at home is so early that the school must have ended before 1 p.m. The arrivals at home slightly before 1 p.m. indicate that the school ended after the fifth lesson around 12:15. This indicates that a person is using the scooters on these two weekdays regularly, to have a convenient possibility to get home. On other weekdays with regular school endings, a school bus might be used for the way home.

In Figure 3b, regular commute trips from the same residential address to a supermarket (in purple) and to the main station (in blue) are visible. The detailed trip information is shown in the appendix in Table 10 and Table 11. The trips in both tables match the usual local working hours, which indicates that these are commute trips. Also it is likely that those trips are done by three individual persons, since there are overlaps in the data that exclude each other: Both the school attendee and the supermarket employee have an overlapping entry on the 25th January, the school kid and the main station commuter have overlapping entries on October 10, November 7 & 9, December 12 and March 5. Also the supposed supermarket employee and the main station commuter have an overlapping entry on November 10 and 22. Given the close proximity of the trip endpoints it can be assumed that all three persons are at least living in the same neighborhood or may even be sharing one household.

A second notable example from the large social group of morning commuters is likely a shift worker of a large warehouse operator. The assumed warehouse employee used the scooters 40 times for their commute, shown in Figure 4. One end of the green trip-cluster is at the warehouse building (Figure 4a), while the other one is next to a house in a region of detached houses (Figure 4b). In one week, shown in Table 5, we were able to find four consecutive days with scooter usage. During this week the user reliably arrived at the warehouse at 9:45 p.m. and at 7 a.m. at home. This matches a working schedule from 10 p.m. until 6.30 a.m. with 8 hours of work plus 30 minutes of break. With multiple weeks with consistent working hours during the week and shifted working hours between the weeks, we were able to reconstruct the working shifts of the warehouse operator, shown in Table 6: a week of morning shifts from 6 a.m. to 2:30 p.m. is followed by a week of afternoon shifts from 2 p.m. to 10:30 p.m., which is followed by a week of night shifts from 10 p.m. to 6:30 a.m. Based on this schedule and assuming that the shifts are fixed, we highlight the expected working hours for days without scooter data in Table 6 in gray. The absence of scooter data may indicate vacation or sick leave, though the person may also have used other means for their commute on these days. Furthermore, it was also possible to find some days they arrived late to work: for example on the Feb 13 2024 the arrival at the warehouse is about 20 minutes later than usual, so they arrived after the regular start of their morning shift.

These data interpretations are illustrative and representative examples that show how it is possible to extract information about different scooter users and their daily/weekly routines, residential and social environment from the sharing platform.

5.5. Validation of the Approach

In order to validate our approach we collected a second, smaller, and independent data set. Over the course of 3 weeks, we collected data in the same way as described in Sections 4.1 and 4.2, again requesting the API in intervals of 3 minutes, for the same town. During these three weeks a group of four a-priori known people used scooters in their daily life. The feature extraction and data processing was done analogously to Sections 4.3 and 4.4. By applying the approach described in Sections 5.1 - 5.3 we were able to gather information about their routines. After the data gathering had been completed, the people were contacted to verify the number of trips and the extracted trip information.

The four individuals made a total of 57 trips, out of which it was possible to identify 52 (91%) of their trips. Thereby, it was possible to gather information about their working routines, sport hobbies, identify visits to the supermarket and personal relationships. A detailed overview

Person	Commute	Supermarket	Hobbies	Visit Friends
A	13 / 0	-	5 / 0	3 / 0
B	-	10 / 1	-	-
C	8 / 0	2 / 2	6 / 0	-
D	5 / 2	-	-	-

TABLE 7: Comparison of the Correct / Wrong Estimated Trips per Person. Grouped by the Purpose of the Trips.

is shown in Table 7. From the five undetected trips, three could not be detected because the start locations were spread too far to be grouped in the same start cluster. The remaining two trips were simply too short, about 300 meters only, so they did not count as trips, following the trip definition in Section 4.4.

6. Dynamic IDs

For the preceding privacy analysis of the scooter users, we used all available data. As mentioned in Section 4 and displayed in Appendix 1, the JSON construct of a scooter contains scooter-specific and static components like the ID, the licence plate and the number code printed on the scooter. These data fields make it easy to identify an individual scooter over an extended period of time.

In fact, the GBFS standard specifies that identifiers such as IDs "must be rotated to a random string after each trip to protect user privacy". Thus, the collected data set is not compliant to GBFS. It is expected that following the standard should lead to less privacy leakage.

Our large real-world data set allows us to evaluate the effect of dynamic IDs on the trackability of users and thereby on the users' privacy. Instead of the ID, which we suppress in the evaluation, we now use the location, current battery level of the vehicle, last location update timestamp, and the last state change timestamp to identify scooters and hence users. We also extended the data set by adding the collection timestamp.

The goal of this analysis is to evaluate how many of the trips can still be reconstructed. Having the complete data set with scooter IDs allows us to evaluate the newly estimated trips against a ground truth. The algorithm is relatively straightforward, with the following steps:

1) Filter out standing scooters, i.e. scooters that did not or hardly (i.e., $\leq$100m) move between two data collections: We obtain a list of appearing and disappearing scooters for each collection timestamp.
2) Identify possible trips candidates: For each disappearing scooter (trip start) store a list of possible appearing scooters (trip end candidates) that match the following requirements:
   - The collection timestamp of the trip end candidate is $\leq$ 2h after the trip start timestamp.
   - The velocity of this trip (distance/time difference) is at most 17 km/h.
   - The battery change of the trip matches the average 2% battery loss per km distance, shown in Table 2.
   - The last state change timestamp of both scooters is equal or the battery was charged (timestamp is during the trip and battery is $\geq$ 100% - 2% per km distance).
3) Estimate trips: Find all trips with only one end candidate and remove other appearances of this end candidate from other trip end lists.
4) Invert the lists: Build the corresponding lists that provide for each trip end a list of possible trip start candidates.
5) Repeat the last two steps until no further changes occur.
6) Return the list with estimated trips that only have one start and one end.

To evaluate the above approach of a privacy analysis under dynamic IDs, a list of ground-truth trips is created by matching disappearing scooters to the first re-appearing scooter with the same ID. The list of the estimated trips is compared to the ground-truth trips and it is checked whether the IDs and collection timestamps of the scooters of each trip start and end are equal to the ground truth.

The results are divided according to different events, as previously specified in Section 4 (trips, re-locations, round-trips, loading, rest). To evaluate the performance of the estimation algorithm, we randomly selected seven days (one for each day of the week) from the recorded 6 months period. The detailed results are shown in Table 8. For example, on October 1 we correctly identified 469 of 578 trips, which is a recall of 81.1%. Over all seven days, we have an average recall of about 80%. Since this is the most important event type for our privacy analysis, we focused in the estimation algorithm on this type, accepting for the others to have a lower recall. In particular, we missed nearly all the loading of scooters, which is a side effect of removing all standing scooters in the first step of the estimation algorithm, since most of the battery swaps happen while the scooters are standing. When comparing the number of total estimations to the number of correctly estimated events over all types, we achieved a precision of 91.2% on the October 1. Both, the achieved recall and precision are high enough to argue that the previously presented privacy analysis, which is based only on trips, also can be performed with a data set that is GBFS compliant.

7. Discussion

In the following we will discuss potential defenses to reduce the privacy leakage in e-scooter systems, which is followed by a discussion of the limitations of our approach. The discussion is completed by an investigation of the generalizability of our analysis to different scooter providers and the ethical considerations we followed during our work.

7.1. Potential Defenses

Countermeasures and potential defenses against such a data collection and subsequent privacy analysis attack can be distinguished into three categories: detection, mitigation, and prevention.

Attack detection requires that the sharing-system provider monitors the API usage and analyzes it for

Date	Weekday	Total estimations	trips	re-locations	round-trips	loading	undefined	recall	precision
Oct 1st 2023	Sun	875	469 / 578	179 / 213	68 / 91	1 / 40	81 / 85	81.1%	91.2%
Oct 24th 2023	Tue	1026	544 / 671	198 / 246	74 / 133	1 / 42	91 / 95	81.1%	88.5%
Nov 15th 2023	Wed	1118	644 / 804	186 / 266	57 / 87	3 / 47	102 / 120	80.1%	88.7%
Dec 18th 2023	Mon	901	494 / 653	125 / 139	57 / 77	8 / 39	94 / 105	75.6%	86.3%
Jan 4th 2024	Thu	687	388 / 493	118 / 128	51 / 60	1 / 3	59 / 64	78.7%	89.8%
Mar 9th 2024	Fr	1003	566 / 694	180 / 209	71 / 104	0 / 22	87 / 96	81.6%	90.1%
Mar 16th 2024	Sat	840	480 / 604	132 / 158	67 / 92	1 / 11	82 / 87	79.5%	90.7%

TABLE 8: Comparison of the Estimated/Ground-Truth Trips under Dynamic IDS.

suspicious activities. Detection of suspicious activity goes hand-in-hand with mitigation techniques to make it harder for an attacker to access the data. Some examples are providing an adequate access control to the API or rate-limiting requests. An adequate access control also allows to track how often a user requests data, which provides the foundation for rate-limiting and user dependent monitoring. However, those mentioned techniques cannot fully prevent the attacker from gaining some access to the data, since also legitimate users have to be able to access the data. An attacker can overcome a rate limit by pretending to be multiple users.

For a more effective prevention of privacy attacks, the available data should be reduced to a minimum. Currently, the API provides clearly more data than the GBFS standard proposes, as can be seen in the listing in Appendix 1. All static or semi-static fields such as the ID, lastStateChange, licence plate, or the scooter code should be removed to prevent a direct identification of a scooter.

We even suggest to go beyond the GBFS definition to reduce the granularity of fields such as the battery level, e.g., by grouping the battery levels into five intervals: $[100, 80)$, $[80, 60)$, $[60, 40)$, $[40, 20)$, $[20, 0]$ (in %). Moreover, associated fields such as the current range meter should be adapted to this granularity. In this manner, the discriminability between scooters is reduced but users still receive sufficient information to choose a scooter. Complementary mechanisms to locally identify a scooter like triggering a signal light or a bell at the scooter are already common practice. Therefore a unique identifier like the licence plate is expendable.

To highlight the improved privacy of publishing only coarse-grained information without unique identifiers, we applied the same analysis of the scooter data as in Section 6 to a data set that only provides the location and a coarse-grained battery level with five possible values, corresponding to the five intervals above. For the data of October 1, this resulted in only 3 out of 578 trips to be identified, a recall of 0.5%. While this seems as if it largely prevents our initial attack, it remains unclear whether more sophisticated attacks, which take the defense into account, are able to reconstruct trips. Thus, we extended our attack through one additional step: After obtaining a set of potential destinations for a trip, we choose the destination such that the average speed to reach this destination from the supposed start point is closest to 10 km/h, the average speed from Table 2. This enhanced algorithm achieved a considerably higher recall, namely 22% or 130 of 578 trips for October 1. Note that we do not claim our improved algorithm to be optimal; another attack could achieve an even higher recall.

Platform	Variable Parameters			Static IDs	
	Battery	Remaining Range	Location	License Plate	QR-code/ID
Tier	%	km	meter	yes	yes
Dott	3 levels	km	meter	yes	no
Zeus	3 levels	km	meter	no	yes
Lime	3 levels	km	meter	3 of 6 digits	no
Bolt	%	km	meter	no	no
Voi	3 levels	3 levels	meter	no	yes

TABLE 9: Comparison of different e-scooter platforms and the granularity of their provided data.

7.2. Limitations

There are several limitations to our privacy investigation of e-scooter platforms. An obvious limitation is that we require frequent usage of e-scooters as our analysis identifies recurring patterns. People who rarely or never use e-scooters are not considered in the analysis. Furthermore, for people that use e-scooters only for solving the last mile problem (or first mile), the significance of the gained information is limited. For example a person, using the scooters on a daily way to a train station does not reveal the final destination of their commute.

Further limitations relate to the empirical character of our study as well as the decisions we have made during the analysis of the data. Obviously, our study is only for a specific town and a specific e-scooter rental company. In particular, for larger cities a scooter rental may only constitute a leg of a total trip and, therefore, privacy leakage may be less. Furthermore, a different population density is likely to affect the number of users and trips, which could make it harder or easier to find patterns. With respect to other e-scooter providers, in our town there was only one company, yet, we analyzed what data can be gathered by other e-scooter platforms (operating in other cities) and present some interesting findings below in Section 7.3.

Next, we discuss some limitations with respect to the way we conducted the data analysis: First of all, we point out that we were not able to track individual people directly, we were only able to observe many independent e-scooter trips. Based on the patterns of trips and the assumptions given in Section 3.2, we conclude that with a high probability some trips are done by the same scooter user. The information that can be extracted by this is always related to visiting a particular place, sometimes also related to the time of visit.

For regions that are often visited by many people and have multiple possible PoIs, such as the city center with many shops, bars, and rental apartments, clustering trips by their location often returns very large clusters that span over multiple streets. Both the large number of scooters in such clusters and the multiple purposes of the groups' region, make data from those regions hardly usable.

Another limitation of our study may be seen in not having ground truth for the most of our data, yet this would have been difficult for ethical considerations (further discussed in Section 7.4). To that end, we provided validation of our analysis for a controlled group of users, see Section 5.5.

7.3. Generalizability

To evaluate the feasibility of our analysis for different e-scooter platforms, we determined which data each provider shares. To that end, we analyzed the mobile apps of the providers and summarized our findings in Table 9. Most providers exhibit static IDs like a license plate or a printed QR-code/ID on a scooter. This helps legit users to identify their rented scooter, but also allows easy long-term tracking of the scooter movements. Only Bolt did not provide this type of static ID. Yet, they provide a combination of scooter locations (within precision of a meter), remaining range (km precision) and battery load (as %-level), thus making it again easily possible to extract most of the trips, using the procedure from Section 6. Only for Bird, we were unable to find which information they provide, which may be due to their bankruptcy in 2023.

This indicates that our method to collect data about scooter usage and extract privacy information should be applicable to other e-scooter providers as well.

7.4. Ethical Considerations

For our work, we periodically collected scooter location data over several months. As we have shown in this paper, this data contains private information like residence, commuting schedules or hobbies of many individuals. In the paper, we mentioned more than 50 individuals as examples but the data holds many more. To protect their privacy, we will not publish the full data set. We carefully selected seven days out of the six month period that cannot be associated directly with users mentioned in the paper. These seven days are used in Section 6 for the analysis of dynamic IDs and potential defenses in Section 7.1. We publish the selected days together with the code used in Sections 6 and 7.1, so interested readers can reconstruct the results that we present in these sections.

Regarding IBR approval, note that our institution has IRB boards on the department level, but not for ours. We followed the ethical guidelines with regard to how we acquire informed consent for our volunteers. The guidelines were not directly applicable for data gathered from what is publicly available on an app, especially since the recorded data is not officially personally identifiable information. We still followed the guidelines as closely as possible by not publishing data that has strong privacy implications and storing it such that only parties involved in the research can access it. In particular, we abstained from live-tracking the scooter data for the whole observation period of 6 months, since this would effectively constitute a stalking of individuals.

To not reduce the privacy of the scooter users further, we decided to not map user profiles to real persons. So we never contacted users that we found via our analysis. Instead, for the validation of our approach in Section 5.5, we had consenting volunteers that agreed to the recording of their data and the subsequent data analysis. As our goal was to show that privacy issues exist at a large scale, relying only on our small set of volunteers for the full analysis would have been insufficient. Still, gathering pseudonymous information about users without consent is a privacy invasion. We considered it acceptable as the data is easily available and could be abused for criminal activities as stalking. In our opinion, the severity of this problem justified our study, which should enable the company to fix or at least mitigate the issue.

In order to conduct a responsible disclosure, we contacted the affected e-scooter company to inform them about the privacy leakage already at an early stage of our work. Since we never received a reaction, we also spoke to a service employee of the company that was interested in our findings and promised to inform his supervisor about this. This also did not lead to any further reaction.

8. Conclusion

Our work reveals that the privacy provided by at least one real-world operational e-scooter sharing system is extremely limited.

For our investigations, we collected scooter data over the course of six month. To that end, we requested the providers API every three minutes for all available scooters in Kaiserslautern, which resulted in 81 million scooter entries with a total volume of 38 GB in JSON files. From this data set we extracted 117k individual scooter trips. By analyzing the extracted trips for spatial and temporal patterns, we were able to cluster trips that were presumably done by the same users. Combining public information from open street maps with the starts and ends of trip clusters facilitated drawing conclusions about the intention of the users' trips. Thereby, it was possible to extract profiles of users containing information about residence, day schedules, and more. We validated our methodology by creating a second, smaller data set. This data set is based on the scooter usage of consenting volunteers in their daily life. Thus, we could apply our approach to this controlled data set and compare the extracted trip information with the trips the volunteers really took.

Furthermore, we analyzed the effect of dynamic IDs, as proposed in the GBFS standard, on our data set and found a simple method to still reconstruct about 80% of the trips. Based on this, we suggested a potential defense by limiting the scooter-individual data that are directly connected to the trips, like the battery level, in order to raise the bar for an attacker.

As indicated in Section 7.1, there are no guarantees that our attack on the supposedly more private data set is optimal. Novel attacks may achieve higher success, rendering our defense also insufficient. As a consequence, it is important to develop methods to publish the data that achieve provable privacy guarantees while at the same time maintaining the utility of data. Differential privacy [17] offers provable guarantees but, as detailed in Section 2, the existing approaches to using differential privacy in the context of mobility cannot be directly applied to shared-mobility systems, which have to enable users to exactly locate vehicles. We hence suggest developing differentially private mechanisms that take the specifics of shared-mobility systems into account as future work.

References

[1] Y. Xu, X. Yan, V. P. Sisiopiku, L. A. Merlin, F. Xing, and X. Zhao, “Micromobility Trip Origin and Destination Inference Using General Bikeshare Feed Specification (GBFS) Data,” *CoRR*, vol. abs/2010.12006, 2020.

[2] G. Baltra, B. Imana, W. Jiang, and A. Korolova, “On the Data Fight Between Cities and Mobility Providers,” *CoRR*, vol. abs/2004.09072, 2020.

[3] H. Li, Z. Yuan, T. Novack, W. Huang, and A. Zipf, “Understanding spatiotemporal trip purposes of urban micro-mobility from the lens of dockless e-scooter sharing,” *Computers, Environment and Urban Systems*, vol. 96, p. 101848, 2022.

[4] J. Mathew, M. Liu, H. Li, S. Seeder, and D. Bullock, “Analysis of e-scooter trips and their temporal usage patterns,” *Ite Journal*, vol. 89, pp. 44–49, 06 2019.

[5] M. Heumann, T. Kraschewski, T. Brauner, L. Tilch, and M. H. Breitner, “A Spatiotemporal Study and Location-Specific Trip Pattern Categorization of Shared E-Scooter Usage,” *Sustainability*, vol. 13, no. 22, 2021.

[6] A. Rubenis, A. Celms, and A. Ratkevičs, “Analysis of Shared Electric Scooter Use in Riga,” *Environmental and Climate Technologies*, vol. 26, pp. 930–940, 10 2022.

[7] A. Hosseinzadeh, A. Karimpour, and R. Kluger, “Factors influencing shared micromobility services: An analysis of e-scooters and bikeshare,” *Transportation Research Part D: Transport and Environment*, vol. 100, p. 103047, 2021.

[8] R. Noland, “Trip patterns and revenue of shared e-scooters in louisville, kentucky,” *Transport Findings*, 04 2019.

[9] S. Bai and J. Jiao, “Dockless E-scooter usage patterns and urban built Environments: A comparison study of Austin, TX, and Minneapolis, MN,” *Travel Behaviour and Society*, vol. 20, pp. 264–272, 2020.

[10] L. Ackermann, M. Mühlhauser, A. Burdusel, M. Federlin, D. Herrmann, S. Holly, D. Nicklas, and D. Wolpert, “Towards Anonymizing Intermodal Mobility Data for Smart Cities,” in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geo-Privacy and Data Utility for Smart Societies*, GeoPrivacy ’23, (New York, NY, USA), p. 24–27, Association for Computing Machinery, 2023.

[11] Y.-A. Montjoye, C. Hidalgo, M. Verleysen, and V. Blondel, “Unique in the Crowd: The Privacy Bounds of Human Mobility,” *Scientific reports*, vol. 3, p. 1376, 03 2013.

[12] J. Freudiger, R. Shokri, and J.-P. Hubaux, “Evaluating the privacy risk of location-based services,” in *Financial Cryptography and Data Security* (G. Danezis, ed.), (Berlin, Heidelberg), pp. 31–46, Springer Berlin Heidelberg, 2012.

[13] D. Kondor, B. Hashemian, Y.-A. de Montjoye, and C. Ratti, “Towards Matching User Mobility Traces in Large-Scale Datasets,” *IEEE Transactions on Big Data*, vol. 6, no. 4, pp. 714–726, 2020.

[14] A. Pham, I. Dacosta, B. Jacot-Guillarmod, K. Huguenin, T. Hajar, F. Tramer, V. Gligor, and J.-P. Hubaux, “PrivateRide: A Privacy-Enhanced Ride-Hailing Service,” *Proceedings on Privacy Enhancing Technologies*, vol. 2017, 04 2017.

[15] Q. Zhao, C. Zuo, G. Pellegrino, and L. Zhiqiang, “Geo-locating Drivers: A Study of Sensitive Data Leakage in Ride-Hailing Services.,” in *NDSS Symposium 2019*, February 2019.

[16] C. Culnane, B. I. P. Rubinstein, and V. Teague, “Stop the Open Data Bus, We Want to Get Off,” *CoRR*, vol. abs/1908.05004, 2019.

[17] C. Dwork, “Differential privacy,” in *International colloquium on automata, languages, and programming*, pp. 1–12, Springer, 2006.

[18] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, “Geo-indistinguishability: Differential privacy for location-based systems,” in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 901–914, 2013.

[19] T. Li, L. Xu, Z. Erkin, and R. L. Lagendijk, “Trajectory hiding and sharing for supply chains with differential privacy,” in *European Symposium on Research in Computer Security*, pp. 297–317, Springer, 2023.

[20] J.-N. Hilgert, M. Lambertz, A. Hakoupian, and A.-M. Mateyna, “A forensic analysis of micromobility solutions,” *Forensic Science International: Digital Investigation*, vol. 38, p. 301137, 2021.

[21] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *kdd*, vol. 96, pp. 226–231, 1996.

[22] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

Appendix A. Further Tables of User Trips

Date	Weekday	Arrival at work	Arrival at home
2023-10-04	Wed	07:24:08	
2023-10-06	Fri	07:34:57*	16:57:51*
2023-11-10	Fri	07:31:55	
2023-11-16	Thu		20:13:22*
2023-11-18	Sat	07:46:39*	
2023-11-21	Tue		16:45:29*
2023-11-22	Wed	07:40:22*	
2023-11-23	Thu	07:53:31*	
2023-11-24	Fri	07:38:09	
2023-12-05	Tue		16:56:42*
2023-12-28	Thu	07:49:07*	
2024-01-05	Fri	07:30:58	
2024-01-11	Thu	07:39:09	
2024-01-13	Sat	07:37:00	
2024-01-24	Wed	07:54:55*	16:48:14*
2024-01-25	Thu		16:52:15*
2024-01-26	Fri	07:45:38*	
2024-02-06	Tue		15:59:13
2024-02-12	Mon		20:11:40*
2024-02-21	Wed		20:12:49*
2024-02-27	Tue		20:18:23*
2024-03-07	Thu		17:23:32*

TABLE 10: Commute of a supermarket employee. (* = back entrance of the supermarket.)

Date	Weekday	Arrival at work	Arrival at home
2023-10-10	Tue	05:54:17	
2023-10-11	Wed	05:58:01	
2023-10-12	Thu	05:50:32	
2023-10-23	Mon	05:50:45	
2023-10-27	Fri	05:53:50	
2023-10-31	Tue	05:50:28	17:34:45
2023-11-03	Fri	05:48:29	17:28:41
2023-11-07	Tue	05:52:59	
2023-11-08	Wed	05:58:00	
2023-11-09	Thu	05:52:08	
2023-11-10	Fri	05:59:46	
2023-11-13	Mon	05:52:12	
2023-11-17	Fri	05:56:55	
2023-11-20	Mon	06:05:22	
2023-11-22	Wed	06:00:36	17:31:28
2023-11-27	Mon	05:51:22	
2023-11-29	Wed	05:52:39	
2023-12-01	Fri	05:56:10	17:27:22
2023-12-11	Mon	05:52:36	
2023-12-12	Tue	05:50:36	
2023-12-15	Fri	05:53:59	
2023-12-20	Wed	05:57:15	
2024-01-08	Mon	05:48:08	17:35:22
2024-01-12	Fri	06:03:38	
2024-01-29	Mon	05:55:45	
2024-02-05	Mon	05:54:47	
2024-02-12	Mon	05:54:04	
2024-02-13	Tue	05:58:25	
2024-02-23	Fri	05:57:37	
2024-02-28	Wed	05:58:54	
2024-02-29	Thu	06:03:47	
2024-03-05	Tue	05:56:21	
2024-03-06	Wed	06:05:27	

TABLE 11: Commute to the main station.

Appendix B. Data Availability

We have decided, after careful consideration, to publish neither the complete nor the validation data set because of the severe privacy leakages that we have found in them. For similar reasons, the data collection and live-tracking code is not shared. Instead, we will publish carefully selected parts of the data set and the code to reproduce the findings of Sections 6 and 7.1. See Section 7.4 for more details and our reasoning. The code and selected data will be shared in a public GitHub repository: https://github.com/ErJedermann/They-See-Me-Scootin