Privacy Violations in Riga Open Data Public Transport System

Arturs Lavrenovs
Faculty of Computing
University of Latvia
Riga, Latvia
arturs.lavrenovs@lu.lv

Karlis Podins
Faculty of Computing
University of Latvia
Riga, Latvia
karlis.podins@gmail.com

***Abstract*—Over the recent years public transportation systems around the world have been migrating to digital ticketing solutions. This paper investigates security and privacy aspects of the one such system implemented by Riga municipality called e-talons by analysing published open data containing ride registrations.**

Keywords—differential privacy; public transport; open data; digital ticket; e-talons

I. Introduction

Replacing tried and tested analog systems with digital counterparts not only brings new possibilities and use cases but usually introduces unexpected threats to privacy and security of the system. Public transport ticketing systems is not an exception, issuing valid tickets without payment and extracting all available data from tickets being a few avenues of thought that attract minds of hackers.

A digital ticket system "e-talons" was rolled out in municipal public transport system of Riga, Latvia in 2009. A few attacks on e-talons have been published [1], but this work does not focus on security aspects of the e-talons system but rather on attacks against privacy of e-talons users. There are built-in privacy violations coming from e-talons design decisions, to make privacy matters even worse, on 25.03.2015 Riga municipality came public with an open data initiative and published all e-talons data on 2 randomly selected months (February 2015, January 2016). Published data is anonymized by replacing real ticketID with another ID using a proprietary algorithm. Nevertheless 1 month of data was enough to successfully deanonymize multiple users, i.e. to find all rides users has done within that month. For theoretic analysis of possible attacks, we assume the worst-case scenario, i.e. whole historic ride data is published by the public transport system operator and their identifiers are not changed, i.e. there is single pair (ticketID, publishedTicketID) for any ticketID.

II. Literature Review

Anonymization of data to be published is a difficult problem. Trivial solution is replacing user identifiers (e.g. name, ID number, social security number, phone number, etc.) with pseudorandom numbers (e.g. strong hash with salt). Such approach removes the obvious data items that can be used directly to establish link between a person and a data record. Such approach successfully anonymizes data only in few special cases, e.g. when a data record contains one attribute or few attributes with few predefined and evenly distributed values with large enough dataset. But data can be deanonymized indirectly, using the patterns within the dataset and also using other public datasets.

A well-known example is the deanonymization of Netflix Prize dataset by Narayanan and Shmatikov [2]. Netflix released dataset containing anonymized movie ratings by 500000 users, which was successfully deanonymized using publicly available Internet Movie Database data. Even earlier research in deanonymizing hospital patient records with help of voter registration database was published by Sweeney [3]. This property of deanonymization and differential privacy is formalized in a seminal work by Dwork [4], where it is formally proven that statistical database always threatens not only objects (individuals) within database but even outside the database.

A good overview of available research is given by Narayanan et al. [5]. Montjoye et al. [6] have demonstrated that 95% of users can be uniquely identified by four spatio-temporal data points, while 50% of users can be uniquely identified by just two spatio-temporal data points. The City of New York, USA has released taxi ride information, some data was hashed to protect user privacy, but that measure turned out to be insufficient [7]. It allowed deanonymization of users and revealed potentially embarrassing information about paid tips [8].

III. Data Description and Quality

There are two distinct sources of data considered in this research, first, ride log published by e-talons system owner as part of open data initiative, second, data freely available over NFC from ticket token - containing ticketID and ride history (last 6 rides for some cards).

Public transportation e-talons ride registration data is published by Riga municipality on their open data web site [9]. Data is published for 2 months - January 2016 and February 2015, each day of the month is stored in a separate file. Each file contains column headers and e-talons ride registrations both as comma separated values. Each registered ride contains incremental identifier unique only to single file

978-1-5090-4473-3/16/$31.00 ©2016 IEEE

which possesses no value to our research, garage to which vehicle is assigned, vehicle type (bus, trolleybus or tram), vehicle identification number which is visible to passengers, route long name, route short number which corresponds to route name, direction (e.g. for a route connecting A and B, direction is either A to B or B to A), e-talons unique identifier we refer to as publishedTicketID which is claimed to be encrypted to protect user privacy (we refer to the original unencrypted version as ticketID) and timestamp with second precision. Route is a fixed sequence of transport stops and is determined by combining route short number and direction, route instance is an event that occurred when specific vehicle left its terminus at specific time and arrived at its destination at specific time. Ride is a single instance of registering electronic ticket with NFC reader on board of a vehicle. To our research most importance have publishedTicketID, routes, route instances, rides and timestamps, this data allows to search for movement patterns and attack individual user or user group privacy.

File sizes differ drastically between 2 months which led us to question published data quality. After all the files were combined into 2-month data sets we got 10917654 e-talons registrations for January 2016 but only 5505785 for February 2015, that confirmed our suspicions about data quality as there were no known events that could double public transportation use in Riga in a single year. After reviewing data sets, we identified that February 2015 data set has no trolleybus and tram ride registrations, additional discovered issue was that February 2015 data set is missing e-talons identifiers in 20.5% cases. Because of such issues February 2015 data is unsuitable for attacks on users and we are using mostly 2016 January data set. January data set contains 5026058 bus rides which is about 8.8% less than February, 2516276 tram ride and 3375320 trolleybus rides. The only noticeable issue with January 2016 data set is that it contains data from outside of claimed range, 57773 records from previous month (oldest record 26 days outside of range) and 75 records from next month. This issue is easily explained by poor design of vehicle e-talons system synchronization with central database, and published data is exported by synchronization date not e-talons ride registration date.

Published data does not contain the public transport stop information where users boarded or got off, it is likely that e-talons system does not provide such data because in Riga municipality there are no ticket types that use public transport stop data to calculate ride price and it would require additional unnecessary integration with vehicle systems. Embarkation stop is valuable for deanonymization attacks, and approximate embarkation stop could be calculated by correlating timestamp information with publicly available timetables, where each route has detailed information on planned arrival times at all stops. Of course such statistical approach does not take into account any extraordinary events like unplanned traffic jams, detours, etc.

IV. Establishing Relation Between TicketID and PublishedTicketID

A. Database Lookup

In the worst case scenario, i.e. when all historic ride data is published, linking a physical ticket token to respective publishedTicketID is trivial:

- Retrieve last rides data (set L) from ticket token.
- Search published ride data for publishedTicketID that has L as subset (there can be some time-delay due to historic ride data not being published in real-time by the transport system operator).

Published research [5] shows that 4 data points get 95% success rate, some ticket tokens we have tested store last 6 rides which should have well above the 95% success rate.

B. Deanonymization

Deanonymization is fairly trivial as long as precise movement pattern of the user is known. A willing passenger personA provided to us with minimal information about his movement pattern, that he drives every weekday to work using 2 disclosed public transport routes, providing two 15 minute time frames in which he usually boards the vehicles, such information can be easily gathered by outside observer. We identified only 8 potential matches following and easily eliminated all but one following specified pattern, and discovered all the rides personA has taken within the month.

C. Reverse engineering e-talons Identifier

Most interesting piece of information in published data set is e-talons identifier which is claimed to be encrypted and indeed those do not match any of the real e-talons identifiers we acquired. First thing we look at is all the unique publishedTicketID digit count distribution which is 9 digits - 130673, 10 digits - 185681, 13 digits - 574246, 14 digits - 60351. Such distribution indicates that no hashing or encryption algorithm that provides fixed output length has been used. We analyzed small set of ticketID from most popular types of physical e-talons and identified that personalised and non-personalised plastic e-talons have 10 digit ticketID, e-talons integrated into bank cards also have 10 digit ticketID, Riga resident's cards have 8 digit ticketID and paper e-talons that is used short term have ticketID with prefix 01 followed by 15 digits. Digit count difference in both ticketID and publishedTicketID indicate that encrypted data can be correlated to real. E-talons usage patterns can be observed in Fig.1, 13 and 14 digit publishedTicketID have same short term usage pattern while 9 and 10 digit publishedTicketID follow same long term usage pattern. In the case of short term usage pattern major drop can be observed starting from the first use and spikes can be observed at the amounts short term tickets are sold (e.g. 10, 20) and the amount of passengers that have driven past 20 times with same publishedTicketID is insignificant. Long term usage has slight drop throughout the month, significant number of passengers using e-talons more than 20 times and no spikes indicating limited usage. Therefore, we draw a conclusion that 13 and 14 digit publishedTicketID correspond to paper e-talons, but 9 and 10 digit publishedTicketID correspond to plastic and integrated e-talons which are used in long term.

Observing the Fig. 1 a correlation between 13 and 14 digit publishedTicketID can be observed in the range of [1..20] rides. After 20 rides the number of the 14 digit publishedTicketIDs drops so low that sampling errors are too

influential. In the interval [1..20] rides average ratio between 13 and 14 digit publishedTicketIDs is 9.109 with standard deviation 0.7694. Possible explanation is that 13 and 14 digit represents the same dataset, and factor 9 comes from base10 notation and removal of a leading or trailing zero.

We have identified that within a single month publishedTicketID does not change by picking multiple random publishedTicketID and known publishedTicketID and verifying that their movement patterns do not change. But publishedTicketID are encoded differently between January and February data sets, only 1205 match from intersecting sets of these 2 month publishedTicketID. All of the matching publishedTicketID are 10 digits long and start with 73795, but their movement patterns between 2 months do not match. This artifact can indicate how encoding is being done or ticketID are generated. Moreover, February publishedTicketID are all either 9 or 10 digit long, it is possible that paper e-talons, which ticketID are encoded with more digits in January are the rides missing publishedTicketID in February.

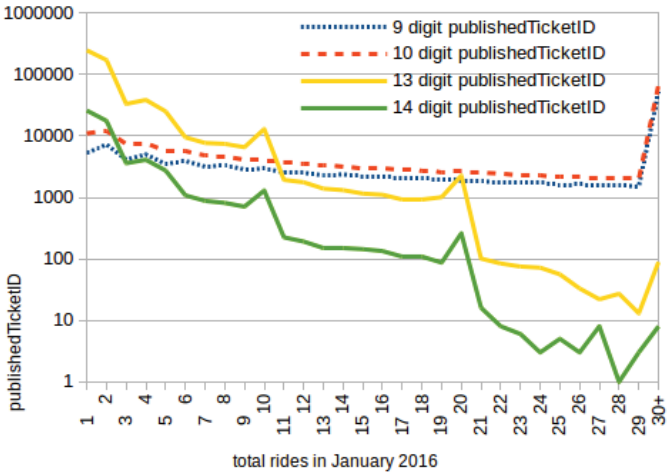

Fig. 1. Ride distribution by publishedTicketID digit count.

We reviewed consecutive publishedTicketID and identified that increment of 3 or multiple of 3 is the most common difference, first 6 multiples of 3 is used as increment differences in 76.8% of cases, only in less than 200 cases there was large increment that was not divisible by 3, possibly separating sets of ticketID assigned for different usage, it indicates that ticketID is not randomly generated and possibly encryption has some multiplication component. Same conclusions can be drawn from publishedTicketID prefixes, 9 digit publishedTicketID has only 10-15 as prefix, 10 digit publishedTicketID start with 6,7 or 9 which means publishedTicketID are closely grouped together. There are no publishedTicketID with the same digit count as the Riga resident's card, it furthermore suggests that multiplication component has been used. This analysis confirms that no secure hashing or encryption algorithm has been used to generate publishedTicketID, most likely scenario is that one or more simple mathematical operations were used. To reverse engineer encryption algorithm, we used only 2 pairs of ticketID and publishedTicketID that are known to be correctly matched. Described analysis and most significant digit difference between chosen pairs instantly indicate multiplication with 3, but this single operation did not produce publishedTicketID from known ticketID, by comparing differences between ticketID*3 and publishedTicketID both pairs produced same result -3072913, thus from these 2 pairs the encryption algorithm is:

$$publishedTicketID = ticketID * 3 - 3072913$$

We verified with owners of more than 10 previously unmatched publishedTicketID and are certain that this encryption algorithm that turned out to be 2 simple mathematical operations was used for all the published January rides.

D. *Identifying Embarkation Stops*

Published data does not contain public transport stops where passengers get on or off, it is more likely that this information is not present at all in the e-talons system than it has been removed because of privacy concerns. Passengers have to register their ticket as they get on board, so timestamp should be approximately that of a vehicle stopping at the public transport stop. No actions involving e-talons are required for disembarking, so there is no information available regarding disembarkation stop (passengers usually go at least one stop, but disembarkation stop could be any stop within set [next stop, …, last stop]).

Information about utilized public transportation stops is most useful for attacking user privacy as it allows to identify work and living area, hobbies and participation in public events. Published data contain vehicle identification number and ride registration timestamps which are sufficient to allow identification of most stops. Vehicle identification number together with route short name and direction is used to identify individual route instance start. We determine that route instance starts when first ride is registered after long period of time or ride is registered after route short number changes, or direction changes.

Public transport routes and schedule is public information published on Riga municipality transportation company web site. By mapping public schedule with e-talons registration timestamps we can group passengers together into likely stop they got on. When data indicates that vehicle has started its route instance we match it together with the route scheduled to depart from terminus. This approach has issues caused by traffic jams, road repairs, route changes, skipped or interrupted routes caused by broken vehicles, etc., but it works best when there is no other competing traffic on the roads and departure and finish time is matching perfectly against schedule.

Most popular bus route and overall second most popular public transport route is A3, our victim personA is also using this route as part of his daily routine. Single morning pre-rush hour route instance histogram with minute interval is displayed in Fig. 2, difference between route instance start, finish and scheduled times is only 1 minute. Even with these conditions that are close to perfect it is impossible to precisely map all the stops to passengers from single route instance, in presented case single ride can be mapped to about possible 3 stops, this number is dependent on route planning, the bigger

distance between stops the more precise this method can be. We can determine exact stop by intersecting possible stop sets from multiple days if victim has pattern of using selected route or alternative routes from the same stop. We applied this method and identified personA exact boarding stop from sampling additional 2 random routine mornings and mapping those against route schedule.

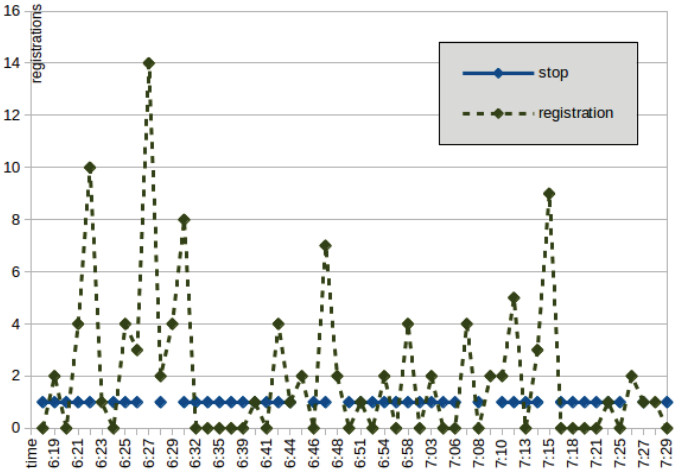

Fig. 2. Single A3 route instance ride registrations.

Another approach is using ride registration spikes which occur when passengers enter the vehicle. Spikes can be grouped into number of groups that equals the number of the route stops. It eliminates all the described issues with schedule matching in fact it does not use scheduling data at all, the only required data is route stops. This approach relies on the fact that most passengers try to register their e-talons as soon as they enter the vehicle, thus borders can be drawn before each registration spike. The unexpected registration spikes could be attributed to ticket control when free riding passengers rush to register their ride to avoid penalty. This approach works well for rush hour routes and routes that pick up one or more passengers at each stop, which is not the case for the few first or last routes of the day and also some midday routes.

Most popular public transport route is 6th tram, 2 instances of this route is displayed in Fig. 3 using registration spike method, first instance outside of rush hour with very few passengers departing at 5:32 and second in the middle of rush hour departing 8:21. First route instance has less registration spikes than scheduled stops which makes it hard to determine precise user stops because it is unclear which stops had no passengers boarding the vehicle. Second route instance has more spikes than stops and it can be assumed that all stops have boarding passengers, multiple consecutive spikes can represent single stop which can be correlated together by applying rule developed specific for route, e.g. expanding spike window, after which stop information from rush hour routes is more precise than scheduled approach.

Both described approaches have problems, combining them together into hybrid approach could solve precision problems and it can be improved even further by combining with additional information, e.g. stop popularity from population density.

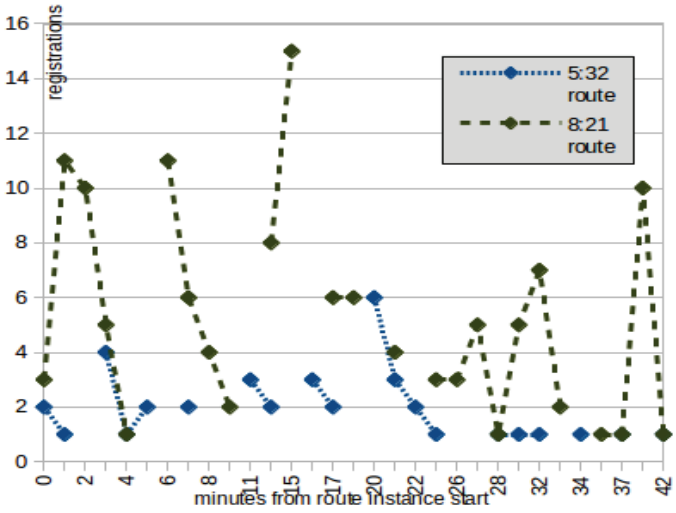

Fig. 3. 2 route instance ride registration spikes.

V. Attacks on Privacy

Both e-talons system design and published usage data violate privacy of e-talons users. Combining both data sources enables additional attacks on privacy of e-talons users. Discussion is split in two subchapters. Simple attacks chapter focuses on exploitation of published ride data. Combined attacks discuss scenarios where published ride data is used in combination with other data sources, either readily available online or where real-life interaction is necessary.

A. Simple Attacks

The heading of this subchapter is somewhat misleading and does not refer to complexity of the attacks. We will discuss attacks with single data source - the ride data published via the open data initiative. Privacy research is especially focused on information leaks in sensitive areas - political, religious, sexual preferences being the most valuable to the individual.

If the ride data is published on consistent basis, it increases possibilities to extract sensitive data about individual ticketID holders. Considering that pricing policy and convenience encourages the use of personal tickets, high number-of rides or bank-card e-talons combos, it is safe to assume that there will be considerable amount of individuals using a ticketID for prolonged periods of time. Timing and location-based patterns disclosing user's political, sexual and religious preferences do exist. We will analyse some examples in the following subchapters.

1) Patterns for religious beliefs

Different religions perform their weekly services on certain days of the week (e.g. Muslims on Friday, Christians on Sunday). Similarly, annual religious festivals are held on distinct dates (Ramadan for Muslims, Easter for Christians). As for location pattern, locations of religious buildings or other properties where religious acts, festivals, etc. are performed are publicly available.

2) Patterns for political opinions

Political parties have both regular weekly or monthly

meetings and unique events like congresses, conferences, etc. The locations of those events are also available. Also protests like marches, rallies and demonstrations have publicly available timing and location information.

3) Patterns for sexual orientation and preferences

LGBT events (e.g. pride), meeting places, red light districts are a few examples for location and timing data that could indicate sexual preferences that users might want to keep private. While this is a lesser threat in small city like Riga, Latvia, some large cities have specific neighbourhoods that host abundance of entertainment venues corresponding to specific preferences.

4) Time and location based attacks

We have provided several examples where sensitive and private preferences correlate with time and location data. The generic attack is to determine whether a ticketID user is member of some particular community. This goal is reached by observing anomalies in rides corresponding to a given ticketID and correlating them with events of that particular community. Since disembarkation stop is not known, identifying rides to a particular event is more error-prone instead commuting back from a particular event should be used if possible. Determining membership in particular community is not possible with attendance of single event due to high amount of noise. But accumulating and analysing data over a long period of time should enable determination of community membership.

B. Combined attacks

This subchapter discusses attack scenarios where published ride data is combined with other data sources. Other data can be already published online or available via some real-world interaction with the target. The generic attack pattern involves several steps (attack can start with step 2 and skip step 1, as for timing-based attacks described below):

- Determining ticketID the targeted individual is using.
- Determining corresponding publishedTicketID.
- Analysing ride history for publishedTicketID.

1) Interrogation of electronic ticket

Electronic ticket communicates privacy-degrading data to any device supporting NFC standard. An android smartphone with NFC reader and publicly available android app is all one needs to retrieve ticketID and last 6 rides. It can be assumed that ticketID and several last rides is public information for any given person. There are several ways how to obtain this data:

- Get in close proximity and run NFC communication session on a standard device:
  - Putting a phone close to a pocket or wallet where ticket token is stored might be enough, depending on circumstances.
  - Attacker can follow the target user, pretend to be a ticket control officer and ask for the ticket.
  - An insider attack scenario (e.g. suspicious partner, employer, colleague) where physical access is easily available.
- As with any other type of radio communication, the distance is limited by the output power of NFC reader and antenna size to receive data from e-talons ticket. So a remote attack from long distance is possible.

Interrogation attacks could be detected by careful observer, but it is not very likely. To defend against this attack, authentication between ticket and NFC reader should be implemented. Probably that is too cost-intensive to be considered, because more expensive ticket tokens need to be used and key management on a distributed system might be challenging. The only feasible defence that users can deploy is usage of RFID-secure storage containers for their ticket tokens at all times, but this does not protect against fake ticket controller attack described above.

2) Timing-based attacks

While interrogation attack could theoretically be detected, timing based attacks are totally passive and undetectable. Several timing-based scenarios can be envisioned:

- Target user is put under physical surveillance, followed on board of vehicle and exact time of ticket registration is recorded.
- Public domain images from some political PR events could contain enough information in photo metadata and visual content. Timestamp and geographical location could be recovered from metadata, vehicle ID and route number could be extracted by looking at the images.
- Similar to public domain images attack scenario, chatty social network accounts could be targeted. In a twitter scenario public transport data is revealed in the tweet text (e.g. “bus X so overcrowded again!”), while timing and geographical location data is available as tweet metadata.

A single instance of public transport usage retrieved by any method described above might not be enough due to large amount of other passengers registering their tickets in the same timeframe. But a set of few rides should be able to identify the corresponding publishedTicketID in the published ride data. Such methods effectively defeat any time-constant algorithm linking ticketID and publishedTicketID, but a strong hash algorithm with daily salt is resistant to this class of attacks.

3) Analysing ride history

There are several data items to be extracted from a ride history for a given publishedTicketID. Conceptually they can be distinguished between pattern search and anomaly search. Once patterns are identified, anomalies, i.e. rides that do not fall into established patterns can be determined. It is not possible to extract any interesting data from infrequent users, a few rides over several years are just not enough, but a frequent user discloses a lot of interesting information.

a) Pattern search

Strong daily usage patterns reveal locations of home, work and places of study. Naturally, people leave from a stop close by to their home, get off close to their place of work or study, just to do the same in reverse in the afternoon. Taking into account the embarkation stop determination algorithm proposed in chapter IV, selecting the most frequently used morning stop as users’ home stop is a safe bet. Following the same way of reasoning, most frequently used afternoon stop as

work or school stop a reasonable assumption. Home, work and school stops are likely to be in reasonable walking distance from the respective stops, unless another means of transport is used, which does not use the same ticketing system (long distance bus, train, park-and-ride, etc.).

b) Anomaly search

Once normal behaviour pattern of a user is determined, rides lying outside the pattern can be investigated for any privacy-compromising clues. To give a few examples, rides to or from nightlife neighbourhoods with late rides to office the following day might be interesting to a malicious employer. Identifying discrepancies between ride data and spoken word might lead a partner to a suspicion of having an affair.

C. Cross-sector attacks

An interesting attack to compromise the confidentiality rather than privacy is enabled by the e-talons/bank payment card combo issued by at least one Latvian bank. If individual using that card is a frequent public transport user, then validity of bank card (month/year) can be determined by simply adding the default validity period of a bank card to the timestamp in first ride record. We assume that bank card/e-talons combo will be started to use within few days of issuance, we estimate that about 1 month is a safe bet. This demonstrates that coupling several systems together increases risks, sensitive financial data is compromised through e-talons ticketing system

VI. Recommendations for Enhancing Privacy

As we have demonstrated, there are several generic attack methods that can be applied by the attackers for multitude of motivations. Suspicious partner, employer or malicious government looking for opposition supporters are just a few of possible attackers, each with their own motivation and resources. Recommendations are divided between those users can deploy and those to be implemented by e-talons system owner.

A. Recommendations for Users

Passengers could use RFID-proof containers for storing ticket tokens or use disposable tickets for 1 ride only. This brings penalty of being excluded from both volume discounting (buying 10 ride ticket is some 10% discount) and limited time tickets (day, week, month). Users should avoid using bank card/e-talons combo. Particularly worrying is that social groups that have discounted or free rides must give up their privacy to be able to receive the benefits (elderly persons have free usage of public transport, but they have a personalised ticket token that needs to be registered on each ride, risking a fine for not failing to register their ticket). Using a 1 ride throw-away ticket for rides user wants to keep private is a feasible strategy too.

B. Recommendations for System Owner

E-talons owner should reconsider length of ride history data stored on the ticket. Regarding publishing the ride data, each day publishedTicketID should be refreshed, for example by using secure hashing algorithm and daily salt, i.e. publishedTicketID=hash(ticketID, dailySalt). This would effectively disrupt majority of attacks described above, where published ride history is used while still enabling mass transit scientists to study usage patterns in public transportation. An evil partner attack reading last 6 rides from ticket token is still possible. Being able to track any user over 1 day should be good enough for studying movement of passengers, e.g. to optimize public transport network, while severely limiting possibilities for attackers.

VII. Conclusions and Future Work

Presented research demonstrates that currently used way of publishing open data containing public transport rides violates privacy of public transport users in Riga. Claimed protection of user privacy by encryption turned out to be only combination of 2 simple mathematical operations and allowed us to gain data set of real e-talons ticketID for all the registered rides of January 2016. We have provided recommendations for system owner to fix this issue. We have demonstrated deanonymization and boarding stop identification attacks, theorised about multiple other attack types. Results also raise question about security of the original stored data which contains much more personal information.

We are particularly awaiting publication of open data containing other months in which some significant political or sensitive events occurred to demonstrate identification of event participants. It is likely that ticketID has significant digits that allows to determine specific type of e-talons used even when digit count match, e.g. schoolchildren versus bank card user, such analysis requires large data set of ticketID and corresponding ticket type and is planned as future work.

References

[1] J. Jansons, Drošības problēma Rīgas Satiksmes biļešu sistēmā "E-talons" (2014), http://possible.lv/news/drosibas-problema-rigas-satiksmes-bilesu-sistema-e-talons/.

[2] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets", 2008 IEEE Symposium on Security and Privacy (sp 2008), Oakland, CA, 2008, pp. 111-125.

[3] L. Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality", The Journal of Law, Medicine & Ethics, vol. 25 (1997), pp. 98-110.

[4] C. Dwork, "Differential privacy", In Proceedings of the 33rd international conference on Automata, Languages and Programming, vol. 2 (2006), pp. 1-12.

[5] A. Narayanan and E.W. Felten, "No silver bullet: De-identification still doesn't work." White Paper, Jul (2014).

[6] Y.A. Montjoye, C.A. Hidalgo, M. Verleysen and V.D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility", Scientific Reports vol. 3 (2013).

[7] V. Pandurangan, On Taxis and Rainbows: Lessons from NYC's improperly anonymized taxi logs (2014), https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1.

[8] J.K. Trotter, Public NYC Taxicab Database Lets You See How Celebrities Tip (2014), https://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546.

[9] Riga municipality, Open data catalogue (2016), https://opendata.riga.lv/satiksme