Building an Open Personal Trajectory Repository

Ville Mäkinen, Anna Brauer, Juha Oksanen
Department of Geoinformatics and Cartography
Finnish Geospatial Research Institute FGI, National Land Survey of Finland
Otaniemi, Espoo, Finland
email: ville.p.makinen@nls.fi, anna.brauer@nls.fi, juha.oksanen@nls.fi

***Abstract*—Portable Global Navigation Satellite System (GNSS) -enabled devices enable gathering comprehensive personal trajectory data about the movement of citizens. Such data would be extremely useful for example improving the cyclability of cities. On the other hand, the data is personal and can reveal surprisingly large amount of personal information. To develop robust privacy-aware and utility-preserving methods to obfuscate personal trajectory data, we are faced with the problem that no comprehensive data sets exist that is needed to conduct such studies. Therefore, we are designing an open trajectory repository where citizens can donate their data for science. The main challenges are 1) how to motivate citizens to contribute and 2) how can we share the trajectory data back to participants and to society while respecting the participants' privacy.**

Keywords-privacy; microtrajectory; open data.

I. INTRODUCTION

Portable GNSS-enabled devices have become ubiquitous in modern society [1]. In principle, these devices enable gathering accurate location data of citizens. Such data could be used for example to assess the flow of cycling traffic, and thus be used to improve the cyclability of cities [2]. Efforts to mitigate climate change are needed and especially in cities, where transportation accounts for a large amount of greenhouse emissions [3], supporting cycling is an important way to help in this battle.

When used for tracking, smartphones and other GNSS-enabled devices usually record their location every few seconds. This results into accurate personal trajectories (consecutive (x, y, z, t) points) that can reveal surprisingly large amount of information about their carriers [4], especially when linked to other data sets [5]. Sharing such trajectories openly thus compromises the participants' privacy. This has been acknowledged in the EU's General Data Protection Regulation (GDPR). One unfortunate outcome of the regulation is that it has made the access and usage of personal location data practically impossible for anyone other than the companies who collect the data. Usually, only heavily aggregated data is available (e.g., Strava heat maps [6]), whose utility compared to the original data is considerably reduced.

We think that the current situation is unbearable. In the GeoPrivacy project, we are studying methods that would enable the use of personal trajectories while respecting the privacy of their owners. More precisely, we are studying how personal trajectories could be processed so that their utility remains high while their potential to violate owners' privacy is hindered. One problem we are facing is that there are almost no personal trajectory data sets that can be used to conduct such studies. To tackle this, we opted to build an open trajectory donation service, where volunteers can donate their trajectory for science.

The rest of the paper discusses the various issues we have so far faced when designing and building such a service. Section II highlights the previous research relevant to this work. Section III summarizes our previous study about the citizens' motivation to participate. Sections IV and V present the methodology that we have developed for this work and possible data sets that could be produced. The sections VI and VII contain detailed implementation details of the service and potential issues. Finally, section VIII contains the conclusion that we have drawn.

II. PREVIOUS WORK

Anonymization of databases have been studied extensively and numerous privacy protection concepts have been devised, for example k-anonymity [7] and l-diversity [8]. Research has been conducted also specifically with spatio-temporal data, but often the developed methods are applicable for relatively coarse data, such as phone location data from mobile operators [9]. The methods that are applicable to GNSS trajectories often require the data set to be dense. For example, in [10] the trajectories are translated closer to each other to achieve a specific (k, δ)-anonymity. In addition, the presented methods seem to be developed mainly for a fixed set of trajectories. Applying such methods to constantly growing data sets may pose additional issues.

There are some studies about anonymization of individual trajectories. For example, truncating the trajectories from their endpoints has been studied in [11]. However, the method does not consider the environment of the trajectory in any way, which in turn can lead to unnecessary utility loss.

As mentioned, open personal GNSS trajectory data sets are scarce. The only publicly available data set we are aware of is the GeoLife data set [12] which contains ~17000 trajectories recorded by 178 individuals in Beijing, China.

Copyright (c) IARIA, 2022. ISBN: 978-1-61208-983-6

We are aware of the Bike Data Project [13], in which cycling data has been gathered. However, they seem to publish only aggregations of the donated trajectories.

These findings support our hypothesis that more open personal GNSS trajectory data would boost the development of privacy-preserving publication methods of such data.

III. How to motivate participants?

The obvious first question is: is there anybody who might be interested in donating their movement data, and what are their motivations to do so? We conducted an online questionnaire to map citizens' thoughts and opinions on the topic. The survey was targeted to Finnish citizens. The results of the survey suggest that the general attitude towards donating personal trajectory data to an open data repository is positive [14]. At the same time, the participants seemed to be aware of the privacy issues. The most important motivators seemed to be the possibility to help to improve pedestrian and cycling lanes, and to participate in scientific research. These findings gave us confidence to proceed with the plan.

The results also indicate that the most important doubts the participants had about such service are related to revelation of personal information and to what the data will be used for. These tell us that it will be important to build the service in such a way that the users can trust the service.

IV. Privacy-aware trajectory processing

A naive method to anonymize personal trajectory data is to simply replace the subjects' personal information with a pseudo-identifier that is unique to each participant. This approach is valid only if there is no other data that the trajectories can be compared to. With external data, for example the known location of an individual at a certain time, a matching trajectory can be linked to the individual. This in turn may reveal for example the home or the workplace, or any other sensitive location, by inspecting the trajectories with the same pseudo-identifier. Removing the pseudo-identifier may make this harder. However, people's movement behavior, for example the mode of transportation (walking or cycling speed) can help narrow down the possible trajectories that belong to the same individual. Also, people are likely to use the same device/application for recording their movement, which may introduce some identifiable information to the trajectories (average accuracy, systematic errors).

One method to counter these threats is to publish only aggregations of the trajectories. An example of this approach are the heat maps by Strava [6]. A more utility-preserving method is to aggregate attributes for example to the segments of the street network of the area [2]. These and similar methods are the most privacy-respecting ones. At the same time, they reduce the utility of the data considerably.

Often the sensitive locations of the trajectories are the endpoints and stay points, i.e., locations where the trajectory spends a considerable amount of time. These can be at the endpoints of the trajectory (user forgets to turn the tracking off after arriving at the destination) or somewhere in the middle (stop at red lights, a quick stop at a shop). We have studied a method to truncate trajectories in such a way that sensitive locations (endpoints and stay points) can be narrowed down only to a group of locations nearby the real location [15]. This approach mitigates some of the threats but leaves (possibly most) parts of the trajectories untouched.

V. Possible open data products

A. Aggregated data sets

Aggregation is commonly used and somewhat "safe" method to publish sensitive data. We have considered to publish two different aggregated data sets. The first one is an aggregation to a regular grid with coarse enough grid size. The second one consists of deriving various attributes from the trajectories and aggregating them onto the segments of the underlying street network, like in [2].

B. Obfuscated trajectories

Publishing even parts of the raw trajectories must be done with extreme care, as the data is personal. We are planning to use the method in [15] as a basis, but further process the trajectories to obfuscate for the exact temporal information. In addition, we have considered resampling the trajectories to a uniform time resolution to make distinguishing them from one another more difficult. This list is not exhaustive and studying and developing new methods in this context is at the focus of our research.

VI. Implementation details

We aim to keep the implementation of the service simple. It will consist of a Vue.js frontend and a Django backend. In addition, there will be an off-line data storage where the original trajectories are stored (Figure 1). The most interesting and important questions are related to processing the trajectory data and management of the possible open data products.

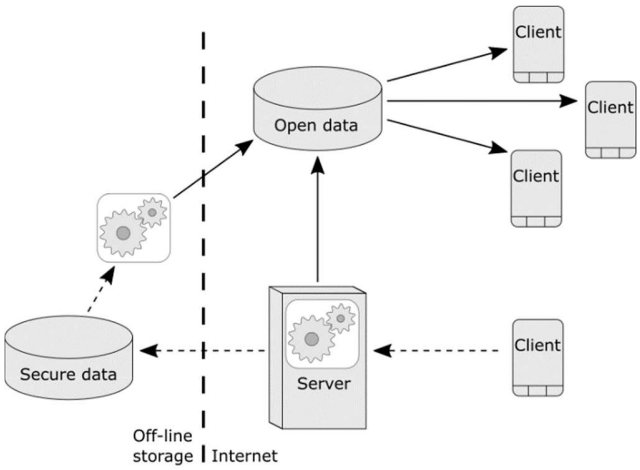

Figure 1. Schematic representation of the architechture of the service. The arrows indicate the data flow (dash dotted: original data, solid: anonymized data).

We argued above that data aggregation is the safest way to publish sensitive spatial data. However, when new trajectories are donated (or some existing ones deleted), we must have all the previous data available to recalculate the aggregations. This is not a problem here since the motivation for us to set up the service is to collect trajectory data. However, we do not want to store the original data on the server that is connected to the internet. A more responsible way is to store the original trajectories on the previously mentioned off-line storage and upload recalculated aggregations to the server when needed. In this scenario, the raw data exists on the backend only for a moment when data is donated but before it has been transferred to the secure storage.

Another option we have mentioned is to publish obfuscated individual trajectories without any further aggregation. In this scenario there is no need to store the original data because there is no need to recalculate any aggregations. It would be possible to perform the trajectory obfuscation on the client-side, and the user could see the results and decide trajectory-wise whether they are happy with the results before donating. In a "production" version of the service this is an interesting approach, because it reduces the role of the trusted party that now stores the original sensitive data. In the first phase of the implementation, we will have a hybrid version of this idea: the data will be uploaded to the backend for processing, but the results are sent back to the user for inspection, and only after explicitly consenting are the results stored permanently in the service.

It should be noted that if we want the participants to be able to request to remove their data, there must be a way to link the obfuscated trajectories to their accounts. Using a pseudo-identifier is the simplest option. More privacy-conserving way is to use a unique identifier for each trajectory and an index that is not public.

VII. IDENTIFIED ISSUES

Aggregated data products have the least privacy issues. Still, care must be taken so that no individual trajectories can be seen or deduced from the data. This can be achieved by filtering out grid cells or street segments that contain less than a predefined number of trajectories.

The obfuscated raw trajectories are more problematic. It is impossible to obfuscate the trajectories in a bulletproof manner, because it is always possible to devise another more comprehensive external data set that the trajectories can be linked against to reveal some personal information. The underlying question is: is a trajectory that is obfuscated in a certain way still personal data? Currently there does not seem to be a clear answer for this.

Originally our plan was to make releases of the database that could be stored indefinitely and used as a benchmark data sets in scientific research. However, the GDPR's right to be forgotten raises concerns whether this is feasible.

The service we have described here is designed mainly for collecting data for research. Even though one of the motivators that was identified in the questionnaire was the ability to improve pedestrian and cycling lanes, we are aware that we are not likely to be able to affect the city planning during this project.

We are also facing the chicken and the egg problem. One of the goals for setting up the service is to gather personal trajectory data that can be used to develop more robust obfuscation methods. At the same time, robust and well-tested obfuscation methods would increase the trust to the service and possibly attract more users.

One big reason for the success and popularity of the existing tracking applications is the ease of use. Having to use a separate service to manually upload data will inevitably drive away potential motivated users. Having a dedicated app for our service might attract more users and should be kept in mind for further development. However, for a relatively small research project such things are big investments. In the first phase we will concentrate on making the user experience of the service as smooth as possible without compromising security.

It might be possible for the user to allow the service to access a third-party service and download data automatically from there. Questions that rise immediately are a) is such use in accordance with the terms of use of the third-party service and b) is it possible to download the original trajectory data programmatically. Usually, it is possible at least for users to download all their data from a service. Developing the import functionality of such data exports from a few popular services may be more cost-effective, given that the trajectory data is provided in a format that can be read for example by the common open-source libraries.

VIII. CONCLUSIONS

The higher goal of the project is to advance the use of ever-growing personal location data for the common good without compromising the privacy of the individuals. The results from the questionnaire suggest that there is willingness in general to participate in such an endeavor. Acquiring the trust of potential participants is of paramount importance. We will try to accomplish that by being 100% transparent with the goals of the service, the methods used, and the possible privacy threats. In practice, these must be communicated to the participants as clearly as possible, in plain language, for example when we ask for their consent to donate data.

We acknowledge that it is entirely possible that there is no "universal" privacy-preserving obfuscation method for personal trajectories, but instead a suitable aggregation must be done case by case. Even in that case, the different aggregations must be generated and published with a thought. Generating and delivering aggregations carelessly may lead into a situation where different aggregations can be combined to reveal more information that what was originally intended.

Copyright (c) IARIA, 2022. ISBN: 978-1-61208-983-6

The results from the research in the GeoPrivacy project will provide more insight into this area as well.

We consider the service that we are building as a test case of a more official service that would be maintained for example by city officials. That would enable a natural connection between the donation service and city planning, make it easier to provide feedback back to the citizens and thus motivate more people to participate.

ACKNOWLEDGMENT

The project is funded by the Finnish Cultural Foundation. We made use of geocomputing platform provided by the Open Geospatial Information Infrastructure for Research (Geoportti, urn:nbn:fi:research-infras-2016072513) funded by the Academy of Finland, CSC – IT Center for Science, and other Geoportti consortium members.

REFERENCES

[1] N. Bento, "Calling for change? Innovation, diffusion, and the energy impacts of global mobile telephony," Energy Research & Social Science, pp. 84–100, 2016.

[2] A. Brauer, V. Mäkinen, and J. Oksanen, "Characterizing cycling traffic fluency using big mobile activity tracking data," Computers, Environment and Urban Systems, vol. 85, 101553, Jan. 2021, doi:10.1016/j.compenvurbsys.2020.101553.

[3] L. Chapman, "Transport and Climate change: a review", Journal of Transport Geography, vol. 15, pp. 354–367, 2007, doi:10.1016/j.jtrangeo.2006.11.008.

[4] C. Song, Z. Qu, N. Blumm, and A.-B. Barabási, "Limits of Predictability in Human Mobility,", Science, vol. 327, pp. 1018–1021, Feb. 2010, doi:10.1126/science.1177170.

[5] O. Goga et al., "Exploiting Innocuous Activity for Correlating Users Across Sites" in The 22nd International conference on World Wide Web, WWW'13, May 2013, Rio de Janeiro, Brazil. pp.447-458, doi:10.1145/2488388.2488428

[6] Strava, "Global Heatmap", 2022, https://www.strava.com/heatmap [retrieved: 05, 2022].

[7] L. Sweeney, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems vol. 10(5), pp. 557–570, 2002.

[8] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data, vol. 1(1), pp. 3, 2007.

[9] M. Gramaglia, M. Fiore, A. Furno, and R. Stanica, "GLOVE: Towards Privacy-Preserving Publishing of Record-Level-Truthful Mobile Phone Trajectories", ACM/IMS Transactions on Data Science, vol. 2(3), pp. 1–36, 2021.

[10] O. Abul, F. Bonchi, and M. Nanni, "Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases", 2008 IEEE 24th International Conference on Data Engineering, 2008, pp. 376-385.

[11] J. Krumm, "Inference Attacks on Location Tracks". In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds) Pervasive Computing. Pervasive 2007. Lecture Notes in Computer Science, vol 4480. Springer, Berlin, Heidelberg.

[12] Y. Zheng, H. Fu, X. Xie, W.-Y. Ma, and Q. Li, "Geolife GPS trajectory dataset – User Guide", https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/ [retrieved: 05, 2022].

[13] Bike Data Project, https://www.bikedataproject.org/ [retrieved: 05, 2022].

[14] V. Jokinen, V. Mäkinen, A. Brauer, and J. Oksanen, "Would citizens contribute their personal location data to an open database? Preliminary results from a survey", in Basiri, A., Gartner, G., & Huang, H. (Eds.). (2021). LBS 2021: Proceedings of the 16th International Conference on Location Based Services, doi:10.34726/1741.

[15] A. Brauer, A. Forsch, V. Mäkinen, J. Oksanen, and J.-H. Haunert, "My home is my secret: concealing sensitive locations by context-aware trajectory truncation", International Journal of Geographical Information Science, in press.

Copyright (c) IARIA, 2022. ISBN: 978-1-61208-983-6