In a search for silver bullet in anonymisation methods

K. Smolak[1], K. Siła-Nowicka[2], W. Rohm[1]

[1]Wrocław University of Environmental of Life Sciences, Institute of Geodesy and Geoinformatics, 50-375, Wrocław, Poland
Email: kamil.smolak@upwr.edu.pl

[2]The University of Auckland, School of Environment, 1010, Auckland, New Zealand

Abstract

The fine-grained human mobility data, enabling studying movement patterns of individuals, have proven their utility in tackling many problems of the contemporary world. Despite their value, access to these data has raised many concerns regarding privacy, making human mobility studies ethically questionable. These privacy issues are the main force slowing down the progress of human mobility research, making access to large-scale data more difficult. Providing open access to statistically unchanged data and protecting the privacy of individuals are requirements that have to be simultaneously fulfilled in order to ensure unconstrained mobility data sharing. Developed mobility data anonymisation methods are able to meet only one of these requirements, hence the ultimate solution for privacy protection of mobility data has not been found yet. In this paper, we aim to point towards the methodology based on synthetic data generation which promises to be the long-searched silver bullet anonymisation method.

1. Introduction

Spatiotemporal information about mobility of individuals and delivering details on the whereabouts of the majority of the world's populations, is a prominent source of data enabling us to advance our understanding of complex urban systems. These data, harvested through various tracking devices, have been used to provide new insights in studied phenomena, such as disease spread (Knop et al., 2021) and traffic (Barbosa et al., 2018), which were not possible to achieve before.

Although the human mobility research area is witnessing a constant advance, the privacy issues directly related to these studies are the main force slowing the progress. So-called *digital breadcrumbs* left behind by mobile devices can be used to reconstruct the movement trajectory of individuals and infer intimate details of their life or information potentially posing a threat to public security (Fiore et al. 2019). Privacy protection is enforced also by legal frameworks, such as General Data Protection Regulation (GDPR) in which location data are considered as personal data and therefore the best anonymisation measures should be applied on them (European Commision 2016). However, neither a unique methodology nor qualitative metrics for privacy protection have been agreed for mobility data. Lack of regulations impedes access to mobility datasets as their actual owners prefer not to share them to avoid legal complications (de Montjoye, 2018).

On one hand, protecting mobility data is important due to aforementioned issues but on the other hand, if human mobility science is to further advance and become significant in solving problems of humanity, widespread access to data is obligatory. Many promising and sophisticated solutions related to human mobility to be deployed on the full-scale require constant access to complete mobility datasets. Therefore, in the opinion of the authors, finding the solution satisfying both problems, retaining privacy and providing access to large-scale human mobility data is a priority challenge in human mobility science.

2. Advances in mobility data anonymisation techniques

The agreement on the anonymisation framework has not been achieved mostly because no anonymisation method satisfies both sides of the conflict. Methods do not provide a sufficient privacy protection level or if they do, the utility of anonymised data is limited.

Anonymising mobility data is extremely complex due to the high uniqueness of mobility traces. Lowering spatial and temporal resolution and pseudonymization (removing all personal identifiers) were proved not to work because these data can still be used to re-identify individuals (De Montjoye et al. 2013). Therefore, researchers are in the search for sophisticated methods, however, silver bullet anonymisation has not been found yet (Fiore et al., 2019). Optimal anonymisation method would follow the principles of privacy-preserving data publishing (PPDP), that is the datasets would fully protect individuals' privacy. Such data would retain full utility and be freely accessible (Fiore et al. 2019).

Two privacy criteria are considered for candidates fulfilling the requirements of PPDP. These are *k-anonymity* (Sweeney, 2002) and *differential privacy* (Dwork et al., 2006). The former assumes that a person's trajectory is indistinguishable from *k-1* other trajectories in the same database. However, in the majority of proposed anonymisation techniques k-anonymity is reached through spatiotemporal generalisation or other data modification which result in information loss (Fiore et al. 2019). Moreover, it is not clear what level of *k* would ensure privacy protection.

The differential privacy (Dwork et al., 2006) criterion is considered a safer approach. This privacy protection principle has been successful in protecting the privacy of location data, adapted worldwide by companies like Google, to protect their customers' data. Differential privacy criterion is satisfied when the presence of mobility traces of a particular customer cannot be inferred based on the results of analyses. Differential privacy is difficult to apply to mobility data, especially to satisfy the recommendations of PPDP and such a method has not been designed yet (Smolak et al., 2020). At the moment, to meet this criterion, data administrators store datasets in databases and allow access to them only through a set of predefined queries which ensure no privacy breaches. This approach, however, is contrary to the PPDP principles.

In their comment paper, De Montjoye et al. (2018) propose and discuss four approaches to privacy-conscientious use of mobile phone data. These approaches are known as limited release, remote access, question-and-answer, and pre-computed indicators and synthetic data. Currently, most datasets are shared as a limited release, where some sample from a limited harvesting period is shared with a specified group of people under a legal agreement. In this approach, shared data still poses a threat to privacy and is not controlled by the owner anymore, hence it can be stolen or uploaded to the Internet. Remote access and question-and-answer models assume data to be stored on the dataset owner side being accessed directly (remote access) or through a set of predefined queries (question-and-answer approach). When a remote access is established, even for a limited group of people, the data are not protected at all, which combined with the possibility of being accessed by third parties poses a risk of privacy breaches. The question-and-answer approach is a direct realisation of differential privacy, where data are accessed only through a predefined set of queries. This approach does not satisfy PPDP recommendations and provides limited possibilities as information that can be obtained from such a database is strictly defined and usually highly aggregated. The similar issue is related to pre-computed indicators as they provide a limited set of statistics on a generalised level. In the paper of De Montjoye et al. (2018), synthetic data and pre-computed indicators are considered as the same data-sharing strategy. However, we argue that synthetic data should be considered as a separate entity; right now it is the only solution showing a potential to solve the problem of mobility data privacy.

3. Setting an outlook on mobility data privacy protection

Creating a not-exact yet very similar replication of real mobility data using some differentially private mechanism would fulfil the criteria of PPDP (Smolak et al., 2020). If generated data retain most of their utility and at the same time would not release private data, such data could be freely published and explored without limitations. The goal is not to replicate mobility trajectories but to retain individual mobility characteristics and mutual interactions through artificially generated location data embedded in a real spatial environment (Fig. 1).

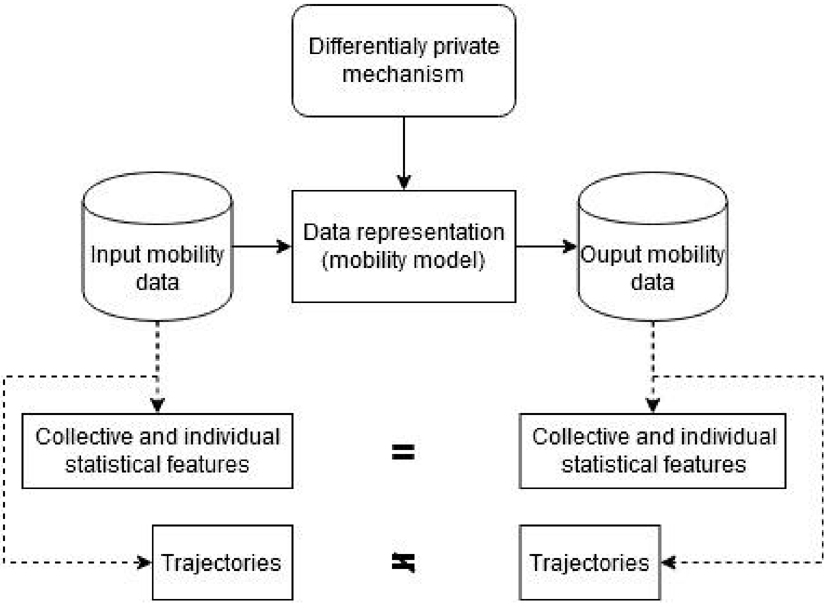
Differentialy private mechanism
Input mobility data
Data representation (mobility model)
Ouput mobility data
Collective and individual statistical features
=
Collective and individual statistical features
Trajectories
≠
Trajectories

Figure 1. A scheme representing the idea of differentially private data generation.

This is, however, an extremely complex task. Few works achieved promising results using an approach where some simplified representations of the original mobility data were used to synthesise artificial individual trajectories (Chen et al., 2012; Mir et al. 2013; Roy et al., 2016). Before data generation, original data representations are modified to meet differential privacy requirements. So far, synthesised data were accurate in a replication of only a few, collective statistics of data, such as hourly population distribution, which normally can be produced out of real mobility data without breaching privacy.

The problem of accurate data replication stems from gaps in understanding general human mobility mechanisms and especially, how individual mobility is related to population flows. However, as human mobility science starts to perceive individual mobility and population flows as a single phenomenon, general mobility mechanisms are becoming much more accurate (Yan et al., 2017). This creates a unique opportunity for the creation of an accurate and universal mobility data replication model.

This challenge is ultimately tied to the evolution of mobility modelling techniques and faces two major problems. First is directly related to the design of the data generation process. An intermediate step of creating data representation is necessary to apply a differentially private mechanism, but significantly decreases modelling accuracy and raises a problem related to spatial embedding of trajectories (Mi et al., 2013). The second issue is related to a tradeoff between privacy and accuracy, because of which a detailed replication will be limited up to some extent. For example, if there is a small group of individuals with a very distinct mobility behaviour, replication of such a group's mobility would potentially cause a privacy breach.

The benefits of deploying mobility data replication technology are worth solving these complex problems. The lack of data accessibility is limiting the development in many areas related to human mobility in academia and industry. Enabling open access to any mobility dataset will attract more research and result in a faster pace of development of many new technologies. Moreover, it will equalise chances to deliver novel products to the market of, as at the moment, mobility datasets are in possession of only a few companies who harvest that data or have direct access to them. If data replication technology would be deployed on a worldwide scale this will ensure full privacy protection of anyone the data are harvested from.

To conclude, the emergence of mobility data has provided valuable insight into the phenomenon of human mobility. However, this has come with a cost of privacy. Currently uptaken measures of privacy protection rely on access restriction, which has an impact on the pace of development of the entire human mobility science and at the same time, does not fully ensure privacy protection (de Montjoye et al., 2018). Providing open access to detailed yet anonymous mobility data is a crucial challenge that should be tackled with priority. In the opinion of the authors, creation of a differentially private modelling model able to synthesise data that well imitates statistical features of the input, may be an ultimate solution for the mobility privacy protection problem.

References

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., ... & Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, *734*, 1-74.

Chen, R., Acs, G., & Castelluccia, C. (2012). Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM conference on Computer and communications security* (pp. 638-649).

De Montjoye, Y. A., Gambs, S., Blondel, V., Canright, G., de Cordes, N., Deletaille, S., … Bengtsson, L. (2018). Comment: On the privacy-conscientious use of mobile phone data. Scientific Data, 5, 1–6. https://doi.org/10.1038/sdata.2018.286

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports. https://doi.org/10.1038/srep01376

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer, Berlin, Heidelberg.

European Commision. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. 59(L 119). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679

Fiore, M., Katsikouli, P., Zavou, E., Cunche, M., Fessant, F., Hello, D. Le, … Stanica, R. (2019). Privacy in trajectory micro-data publishing : a survey. 1–35. Retrieved from http://arxiv.org/abs/1903.12211

Knop, K., Smolak, K., Kasieczka, B., Rohm, W., Smolarczyk, T., & Zyga, M. (2021). Mobility modelling for simulation of spatial spread of infectious diseases. In *EGU General Assembly Conference Abstracts* (pp. EGU21-13112).

Mir, D. J., Isaacman, S., Cáceres, R., Martonosi, M., & Wright, R. N. (2013). Dp-where: Differentially private modeling of human mobility. In *2013 IEEE international conference on big data* (pp. 580-588). IEEE.

Smolak, K., Rohm, W., Knop, K., & Siła-Nowicka, K. (2020). Population mobility modelling for mobility data simulation. *Computers, Environment and Urban Systems*, *84*, 101526.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

Roy, H., Kantarcioglu, M., & Sweeney, L. (2016). Practical differentially private modeling of human movement data. In *IFIP Annual Conference on Data and Applications Security and Privacy* (pp. 170-178). Springer, Cham.

Yan, X. Y., Wang, W. X., Gao, Z. Y., & Lai, Y. C. (2017). Universal model of individual and population mobility on diverse spatial scales. *Nature communications*, *8*(1), 1-9.