DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Crowdsourcing mobility data with privacy preservation through decentralized collection, analysis and storage

Simon van Endern

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Crowdsourcing mobility data with privacy preservation through decentralized collection, analysis and storage

TODO: Titel der Abschlussarbeit auf Deutsch

Author:	Simon van Endern
Supervisor:	Prof. Dr.-Ing. Jörg Ott
Advisor:	Trinh Viet Doan
Submission Date:	30.06.2019

I confirm that this bachelor's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 30.06.2019 Simon van Endern

Acknowledgments

Abstract

We propose a method to publish location ata without raising privacy concerns.

As still this data could be useful for many stakeholders, we will investigate how on the one hand aggregated data can be published without imposing any privacy risk to the owners of the data and on the other hand develop a prototype of a mobile application through which this location data is aggregated in a decentralized manner so that the raw user data never leaves the users' device.

Contents

1 Introduction

1.1 Motivation

1.1.1 General motivation

"Data is the new oil" is an often quoted stigma and means that more and more businesses are based not on specific production capacities but on data, the ability to process it and the exclusive owenership over it. The success and monopoly of companies like Google or Facebook can at least to some extent be attributed to this exclusive ownership.

According to commonly accepted economic theories, monopolies hinder innovation and progress. This implies that the unavailability of huge amounts of data to the public is an impediment of innovation and increased growth.

Some governments and other institutions therefore already publish some of their datasets after anonymizing them. Nevertheless, the applied anonymization is often not sufficient or at least critical. Research shows that inferences can be drawn from the published datasets that violate the respecitve users' privacy. But also privacy concerns of users have increased due to leakages where their data was not well protected at e.g. facebook and stolen and published.

So, we identify two issues compromising data privacy.

1. The availability of huge datasets at central servers imposes a risk stemming from the computer science area of security.

2. Publication of entire datasets can even after applying anonymization techniques not guarantee privacy preservation.

1.1.2 Examples of direct and indirect privacy breaches

An example for the first issue is the facebook data scandal representative for many data breaches over the last years. TODO: [Find and cite].

An example for the second issue is that the location data of Twitter tweets was published without asking the user for permission. Furthermore this data is only available through the API, so that the user is not aware of this infringement. Using this data, [ZB11] has shown that this data can be used to infer a users home address and often also the work address, even if the user itself is privacy-aware, thus does not publish his / her name, etc.

1.1.3 Classification of location data and apps that use it

In order to review existing approaches and research, classify location aware services by the acceptable delay of the location information being available:

- Almost no delation tolerance: e.g. an application showing a pop-up about a nearby venue e.g. a coffe shop when a pedestrian passes
- Some delay e.g. one minute is acceptable: An application e.g. google maps derives the information of congested traffic from devices reporting their GPS data which show lower than usual speed. As congestions worth reporting last longer than one minute, some delay in the device's information reaching the server is acceptable.
- Significant delay of hours, days or even weeks is acceptable for historical and statistical use of location data e.g. to find out about popular visiting times

1.1.4 What has been achieved so far

Most existing approaches focus on publishing location data where a huge delay is acceptable as can be seen in the following table: TODO [create table].

- Collect less data [GP09]
- Mixing approach [BS03]
- Anonymize data to meet the kriteria of k-anonymity [Swe02b] and [Dra+19]
- spatial cloaking [Kru07]
- Remove not only identifiers from the data-set but also apply algorithms, that remove samples, that can be (due to few samples in this area) identified [Hoh+07]

1.1.5 Problems that still arise

Still this privacy is only limited if only this one dataset is taken into account. If e.g. multiple of those data-sets from different data collectors are combined, or information about an individual like home and work adress is provided, privacy breaches are still highly likely. Furthermore, those algorithms always depend on a trusted server to collect the data from all users and then publish the results of any analysis applying privacy-preserving algorithms beforehand. So while all those different approaches to preserving privacy while publishing data-sets manage to achieve ever better results, they always depend on a trusted server for creating the full data-set beforehand. This still imposes a high privacy risk to every user, as trust can either be misued by the trusted server itself or by other parties exploiting eventual security loopwholes in the trusted server.

1.2 Research Question

Thus the two problems stated in 1.1.1 are still widely unresolved and have not been tackled in common so far. We investigate the possibility of storing location data only decentralized on the devices where they it is collected as well as querying this data in a decentralized manner using P2P technology in order to inhibit any instance from accessing raw data.

1.3 Contributions

Thus our approach takes the opposite direction. We do not first collect the whole data-set and then reduce it to a data-set meeting privacy-constraints but we start from the bottom up - first by performing analysis in a decentralized manner so that there never is an overal data-set imposing a security risk on all the entries' users, and second by proposing a framework that only releases aggregated data where no interference of any user information is possible. This data will then be available to everybody. This gives us maximum possible feedback on eventual privacy problems, creates trust through transparency and supports the process of not randomly collecting data and afterwards researching on metrics that are actually needed but first on evaluating which metrices are needed and then retrieving them if possible without raising privacy concerns.

We will use the definition of location privacy as defined by [BS03]: " the ability to prevent other parties from learning one's current or past location". They further propose a different approach to preserve privacy. TODO!!!

We develop a framework ...

Beyond the scope of this research is ...

1.4 Outine

The rest of this research is organized as follows ...

2 Related Work

[Dra+19] finds that even when personal data is anonymized thus that names and addresses, etc. are removed, sensitive information can be inferred from the data. In this study it was shown that from call-records in the US the home address and also often the work address of a person could be inferred. They highlight that while adhering to the k-anomymity model proposed by [Swe02b] it is practically not possible to publish datasets that are still of any significant use.

Also [GP09] highlights the thread that home and work locations can be inferred from anonymized datasets and can in combination with other sources yield even more information about a user. To reduce this risk, they propose "to collect the minimum amount of information needed". In contrary, we want to investigate another approach, so that rich data can still be used and be published in an aggregated manner to let people profit from the data but still preserve privacy.

Another problem that arises is that anonymization algorithms applied to datasets prior to publishing them might yield good results if the location data is in a densily populated area but might perform poorly if the population is only sparse [Hoh+07].

[Hoh+07] identify that while privacy algorithms might successfully provide privacy for location data samples in highly frequented areas, but perform poorly and disclose sensitive information for samples in areas with lower traffic frequency. They discuss the problem commonly accepted in research that either the quality of the data becomes poor or useless when applying techniques like k-anonymity [SS98; Swe02b; Swe02a] or that privacy cannot be guaranteed. They propose a novel algorithm based on time-to-confusion. Thus basically whenever it is possible to attribute two different samples of a dataset with a high probability to the same user, the corresponding sample gets removed from the data-set to be published. This is necessary, as "the degree of privacy risk strongly depends on how long an adversary can follow a vehicle" [Hoh+07]. In more detail, time-to-confusion also takes into account the entropy information provided by the whole dataset, thus that even when two samples cannot be connected with high probability due to to many possible consecutive samples, analyzing the whole dataset can provide information that actually the possible consecutive samples have different probabilities due to common route choices. E.g. a vehicle on a highway is much more likely to follow on the highway for some more time than leaving the highway. While this information is taken into account, they point out the limitations of their work that when the dataset is matched with street maps, even more samples would have to be remoed to ensure privacy because it will render some former possible consecutive samples impossible due to missing streets connecting them.

[BS03] introduces the concept of mix-nodes already known from privacy research on a network level (TODO: "copy" related work part of paper "time-to-confusion"). They propose a framework in which privacy is protected through frequently changing pseudonyms. Furthermore they find that similarly to the problem of identifying consecutive samples in [Hoh+07], the change of pseudonyms has also to be obfuscated in order to provide complete privacy. In contrast, this paper focuses mostly on solving the problem that location aware services that e.g. notify you when you are close to a venue of interest, do not need to have access to your location data at anytime but can register to events with a mix-node. Thus they register for the venues of interested and only get notified when the mix-node, which is trusted and has complete access to location data, detects a match. One sees straight away, that this again depends on trust of the users on the mix-node. Nevertheless, the proposed solution of mix-nodes and mix-zones analyzed on a sample shows that even using this framework, privacy cannot be provided, especially as here again the entropy provided by the history of the released or somehow collected data-set makes it too hard to obfuscate the consecutiveness of different pseudonyms.

[Swe02b] is the current state of the art of minimum data protection. They define a dataset as the commonly understood tables in SQL. Besides the unique identifier used in the table, a quasi-identifier is the combination of several attributes with which a set of entries can be identified. a dataset adheres to the rules of k-anonymity, if querying every possible such identifier returns at lest a set of k different entries. Thus 1-anonymity identifies an entry exactly and provides no anonymity at all. The anonymity problem arises not from the dataset itself, but from a combination of datasets, that have the attributes of the quasi-identifier in common. This way anonymous knowledge from both datasets can be linked in order to infer information not intended to be made public. They also highlight, that also publishing the same dataset with different privacy-rules, i.e. different anonymization techniques applied, can result in inferences that reveal the original dataset.

[Swe02b] clearly highlights that there are two approaches to hiding sensitive information. One is to restrict queries to a database that might reveal sensitive information. In contrast to this approach, they focus on anonymizing the data already before any access to it. Nevertheless, this is based on the assummption that the data owner knows about any possible quasi-identifier in order to obfuscate the dataset sufficiently to provide k-anonymity for all quasi-identifiers. If one quasi-identifier is not thought of, the dataset might expose 1-anonymity for this identifier and result in possible exposures of data not intended to be public.

[Swe02b] also discusses further problems that are easy to tackle but nevertheless necessary to protect users' privacy. The order of the published table must be random. Otherwise there is more information (hidden) available that can be used to break k-anonymity. Another problem is when the same table is released and obfuscated differently for the same quasi-identifier, other attributes in the releases can be used to link entries and thus de-anonymize the data.

[GP09] further investigates the fact that from a dataset containing GPS data of trajectories or e.g. twitter-posts as in [ZB11] the home location can be inferred with high probability. They show that also the work location can be identified with pretty high accuracy and probability. Furthermore they find that people who live and work in different regions or more generally, the further work and home diverge, the smaller the anonymity set of the specific user in the dataset and thus the lower also the anonymity. This is similar to the findings of [BS03] that users in less populated areas are exposed to more privacy risk than in denser areas.

[BS04] extends the analyzis of [BS03].

TODO: Cite middleware usage approach by [GG03] TODO: Cite approach of disclosure algorithms by [GH05] TODO: Cite confusion approach similar to [Hoh+07] by [HG05] TODO: Cite querying an anonymization by [MCA06] TODO: Read [Tan+06]

3 Solution

4 Analysis

5 Conslusion

List of Figures

List of Tables

Bibliography

[BS03] A. R. Beresford and F. Stajano. "Location privacy in pervasive computing." In: *IEEE Pervasive computing* 1 (2003), pp. 46–55.

[BS04] A. R. Beresford and F. Stajano. "Mix zones: User privacy in location-aware services." In: *IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second*. IEEE. 2004, pp. 127–131.

[Dra+19] K. Drakonakis, P. Ilia, S. Ioannidis, and J. Polakis. "Please Forget Where I Was Last Summer: The Privacy Risks of Public Location (Meta)Data." In: *CoRR* abs/1901.00897 (2019). arXiv: 1901.00897.

[GG03] M. Gruteser and D. Grunwald. "Anonymous usage of location-based services through spatial and temporal cloaking." In: *Proceedings of the 1st international conference on Mobile systems, applications and services*. ACM. 2003, pp. 31–42.

[GH05] M. Gruteser and B. Hoh. "On the anonymity of periodic location samples." In: *International Conference on Security in Pervasive Computing*. Springer. 2005, pp. 179–192.

[GP09] P. Golle and K. Partridge. "On the Anonymity of Home/Work Location Pairs." In: *Proceedings of the 7th International Conference on Pervasive Computing*. Pervasive '09. Nara, Japan: Springer-Verlag, 2009, pp. 390–397. ISBN: 978-3-642-01515-1. DOI: 10.1007/978-3-642-01516-8_26.

[HG05] B. Hoh and M. Gruteser. "Protecting location privacy through path confusion." In: *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM'05)*. IEEE. 2005, pp. 194–205.

[Hoh+07] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. "Preserving Privacy in Gps Traces via Uncertainty-aware Path Cloaking." In: *Proceedings of the 14th ACM Conference on Computer and Communications Security*. CCS '07. Alexandria, Virginia, USA: ACM, 2007, pp. 161–171. ISBN: 978-1-59593-703-2. DOI: 10.1145/1315245.1315266.

[Kru07] J. Krumm. "Inference Attacks on Location Tracks." In: *Pervasive Computing*. Ed. by A. LaMarca, M. Langheinrich, and K. N. Truong. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 127–143. ISBN: 978-3-540-72037-9.

[MCA06] M. F. Mokbel, C.-Y. Chow, and W. G. Aref. "The new casper: Query processing for location services without compromising privacy." In: *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment. 2006, pp. 763–774.

[SS98] P. Samarati and L. Sweeney. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and specialization." In: *Proceedings of the IEEE Symposium on*. 1998.

[Swe02a] L. Sweeney. "Achieving k-anonymity privacy protection using generalization and suppression." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 571–588.

[Swe02b] L. Sweeney. "k-anonymity: A model for protecting privacy." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570.

[Tan+06] K. P. Tang, P. Keyani, J. Fogarty, and J. I. Hong. "Putting people in their place: an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications." In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2006, pp. 93–102.

[ZB11] H. Zang and J. Bolot. "Anonymization of location data does not work: A large-scale measurement study." In: *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM. 2011, pp. 145–156.