DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Collecting, analyzing and publishing location / mobility data without raising privacy concerns through decentralized analysis and storage.

Simon van Endern

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Collecting, analyzing and publishing location / mobility data without raising privacy concerns through decentralized analysis and storage.

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.

Contents

1 Introduction

The introduction is meant to motivate the subject area (why is this important?), define the problem you are interested in (what are you doing?), and limit the scope (where do you stop?). It also gives an outline of the thesis (which chapters will explain what?) and explains how you are going the approach your subject.

With the advent of the internet and large-scale applications, the question of privacy has drawn increasing attention. Especially with services like Twitter, Facebook, Google & Co. there are problems and privacy infringements when user data is realeased. One example 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.

[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.

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!!!

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.

This research shows that publishing raw data is critical, even when the data is anonymized. As still this data could be useful for many stakeholders, we will investigate how on the one hand aggregated datacan 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.

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.

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.

They further find that 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. 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.

TODO: Classification of location based services: real time vs. historical. Tracking vs. providing push-notifications of nearby venues.

1.1 Existing approaches

- 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]

1.2 Section

1.2.1 Subsection

See **??**, **??**, **??**, **??**.

2 Related Work

[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.

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.

[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.

[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.

[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.

[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.

[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.