DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Crowdsourcing mobility data with privacy preservation through decentralized collection and analysis

Simon van Endern

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Crowdsourcing mobility data with privacy preservation through decentralized collection and analysis

Crowdsourcing von Mobilitätsdaten ohne Einschränkung der Privatsphäre durch dezentrales Sammeln und Analysieren

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

Contents

1 Introduction

1.1 Why we need an open-source location data approach

"Data is the new oil" is a quote many people agree with. It 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, Facebook and Amazon can be attributed to this exclusive ownership to a significant extend.

While patents that used to power companies' success provide a balance through granting exclusive rights while having to make the knowledge public, many companies e.g. Coca cola have decided successfully not to go for a patent and thus not reveal their knowledge. If that approach is not compromised, it guarantees both - non-disclosure and also exclusive rights. Similarly, the non-disclosure of huge data sets collected by Facebook, Google and Amazon circumvent the balance intended by patents. The unavailability of huge amounts of data to the public is an impediment of innovation and increased growth. For example, cities would benefit from aggregated location data in order to optimize traffic scheduling as also highlighted by [6]. Nevertheless, the publication of raw data sets is impossible because it severly intrudes the privacy of the owners of the data.

So, even if companies would agree on a publication, a problem arises. There is a conflict between preserving user privacy and publishing user data.

Nevertheless, user privacy is already compromised even without publication of user data. Already the mere existence of central data sets pose a privacy risk to users, because security issues might allow for theft and unwanted publication of these data. An example is the theft of 14 million user data from facebook [5].

Some governments and other institutions already publish some of their data sets after anonymizing them e.g. through cloaking of data so that it achieves k-anonymity and there are crowdsourcing and open source approaches to make data available to everybody. Nevertheless, the applied anonymization is often not sufficient or at least critical if the resulting data set should still be useful. Research shows that inferences

can be drawn from the published data sets that violate the respecitve users' privacy. So, in addition to the main risk of a central data set, publishing anonymized data poses another risk to users privacy.

Furthermore, besides the remaining risk of inference attacks in published anonymized data sets, the anonymization through 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 even if the data is only stored anonymized on the server, besides the remaining risk of inference attacks, this still imposes a high privacy risk to every user, as trust can be misued by the trusted server itself.

1.2 Research Question: No central raw data set but only aggregated data

RQ 1: What features does such a system require? RQ2: ... Nach dem Stil.

Clearly, in order to overcome the conflicit between privacy intrusion and (public) data availability, a solution is needed that gets along without storing raw data in a central data set. This solution should 1. eliminate the risk of leaking raw user data through theft from a centralized database and 2. eliminate the remaining risk of inference attacks on published believed-to anonymized raw data. So far, we have not seen an approach to fully solve this problem.

1.3 Contributions

For our solution, we will focus on the sub-area of location data and location privacy. We investigate the possibility of storing raw location data only decentralized on the collecting devices. On a central server available to the public, only aggregated data is stored, thus the main problem of privacy risk by a central database containing the overall raw data set is solved. Furthermore, the issue of trust is removed, as the aggregation process happens decentrally, thus the central server will never hold any other data than aggregated data. It will never know about the individual raw data.

In summary, our approach takes the opposite direction as todays standard. 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 the public. This gives us maximum possible feedback on eventual privacy problems, creates trust through transparency and fosters innovation through availability of data to everyone.

1.4 Outline

The structure of our research is organized as follows: First we re- view related work in the areas of location privacy and anonymisation techniques. In section **??** we describe our approach of decentralized data analysis to get along without a central database. Section XXX describes the setup in detail. Section XXX analysis the result from field-testing our application. Section XXX incorporates the results into our proposal of a possibility to achieve 100% privacy through all applications. Section XXX summarizes our work and points out further research possibilities.

2 Related Work

2.1 Classification of location data and apps that use it

In order to review existing approaches and research, we classify location aware services by the acceptable delay of the location information being available: Such a classification has already been made by [6].

1. 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
2. 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.
3. 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

Most research investigates user's privacy in case 3 [Citations!!!] or 2. For case one there are already solutions available. We will first review research tackling location privaacy in case 3 and 2 and then briefly point out the findings for case 1.

2.2 Research has identified the following privacy problems

2.2.1 Central databases itself pose a risk due to possible theft

Centralized databases also expose the users to a security risk (through theft) [14, 7].

- [9] proposes the use of P2P over WIFI and Bluetooth to decrease the need of central instances.
- [10] proposes a secure approach where the raw data is hidden from the central instance but still the aggregated data can be obtained by using encryption methods. This approach is very close to our work. Also [7] is close to our work and uses encryption.

- [7] proposes an approach to handle user authentication.

2.3 Inference attacks on published data

2.3.1 Inferring home and work location from consecutive data samples

Research has shown, that even from a location data set that is pseudonymous, i.e. the identifiers have been stripped or anonymized from the data, it is still possible to infer the home location of single users through inference attacks [11, 3, 4, 7, 17]. The same problem arises when using data collected through crowdsourcing [10].

2.3.2 Inferring identity from home and work location

Furthermore, this location coordinates can then be combined with publicly available information e.g. reverse map coding of coordinates to addresses and then searching for entries in telephone books to infer the users identity from it's home location [11, 4, 7]. This identity can then be linked to other sensitive data. This problem also arises in the area of IoT [14, 7]. Often (though with usually lower probability) also the work address in addition to the home address can be inferred and makes linking the data to identities even easier [3, 4].

2.3.3 Solutions / Countermeasures to prevent inference attacks

Spatial cloaking: Achieving k-anonymity by dropping data points or perturbing them or dropping all data points around a random point around the home location [11]. More sophisticated approaches: [8]

2.3.4 Problems still after countermeasures

Data suppression algorithms have only limited success and can only reduce, but not eliminate the risk [7].

Data is useless if k-anonymity is guaranteed

Insufficient accuracy / the data set becomes useless [11, 3, 13, 16, 15].

Countermeasures not effective in sparsely populated areas

Anonymization techniques might score well in densly populated areas or areas with high traffic but poorly in sparsely populated areas especially where a single address

can be mapped to a single person or family [8, 1, 7] [location-privacy correct paper or cited wrong paper???] or might not work for individuals whos work and home location are further away than average [4].

Still privacy breaches possible

- More advanced privacy breaking algorithms
- Taking other sources into accouont, e.g. history of location data Extending the time period over which data is collected generally increases the risk.
- quasi-identifiers not thought of

2.3.5 All methods depend on trust to a third party or the provider itself

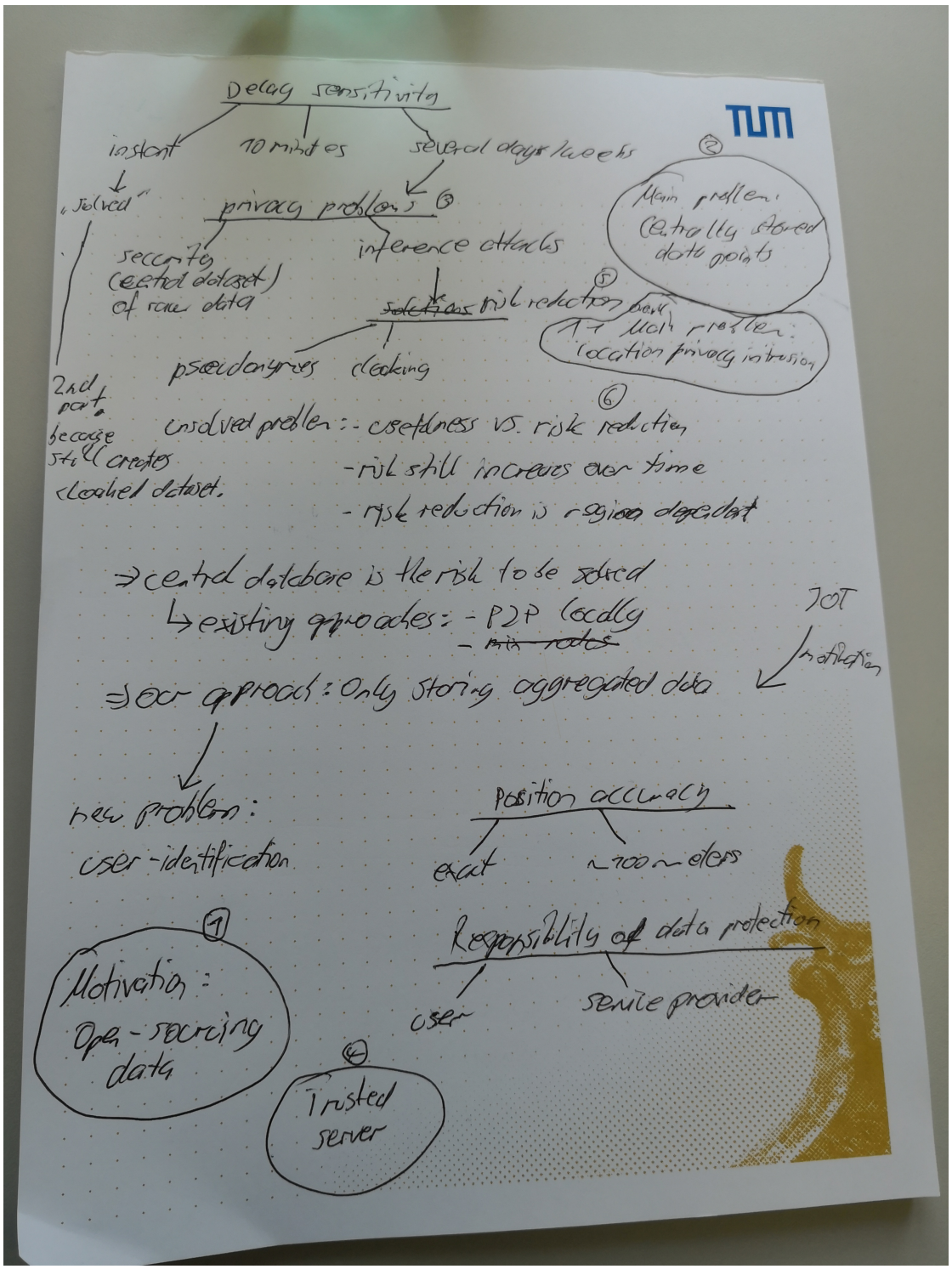
Still all approaches depend on first centrally collecting the original raw data and then before querying [16] applying anonymization techniques.

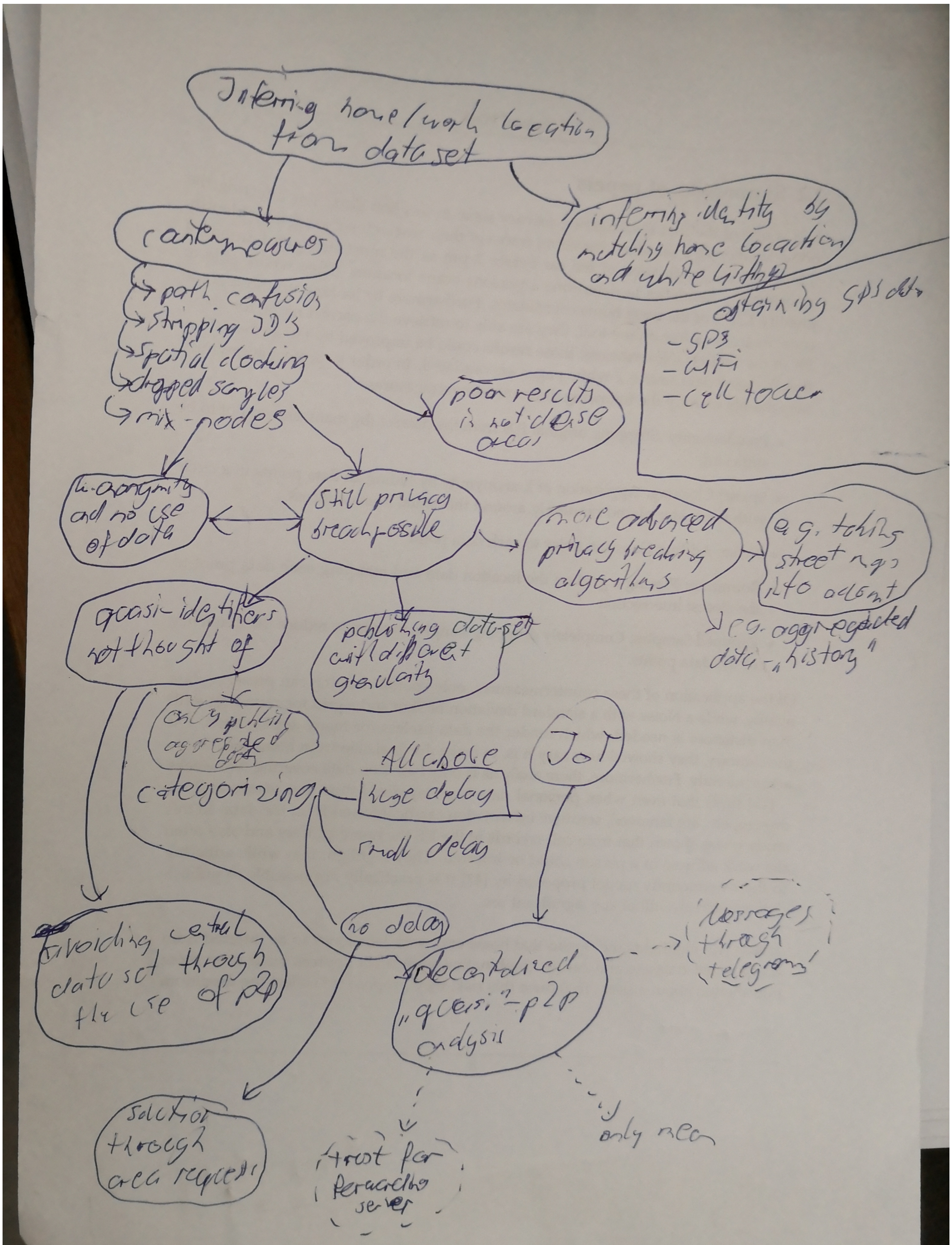
2.4 Category 1 location data use: instant

[1, 2] introduces mix-nodes, that can nevertheless not guarantee privacy and also depnds on a trusted third party. Also [12] proposes a solution (close to our summary) how to enable privacy for instant use of location data.

Decentralized methods for data analysis are also motivated from the area of IoT [14].

TODO: Relate to [16]


Delay sensitivity
instant
10 minutes
several days/weeks
"solved"
privacy problems
inference attacks
security (central dataset) of raw data
risk reduction
pseudonyms
cloaking
2nd part, because still central dataset
unsolved problem: - coarseness vs. risk reduction
- risk still increases over time
- risk reduction is region dependent
⇒ central database is the risk to be solved
↳ existing approaches: - P2P (costly
⇒ our approach: only storing aggregated data
IOT
motivation
new problem:
user-identification
position accuracy
exact
~100 meters
Responsibility of data protection
user
service provider
Motivation: Open-sourcing data
trusted server


Inferring home/work location from dataset
inferring identity by matching home location and white listings
countermeasures
path confusion
stripping ID's
spatial cloaking
dropped samples
mix-nodes
poor results in not dense areas
obtaining GPS data
- GPS
- WiFi
- cell tower
k-anonymity and no use of data
still privacy breach possible
more advanced privacy breaking algorithms
e.g. taking street maps into account
e.g. aggregated data „history"
quasi-identifiers not thought of
publishing dataset with different granularity
only publishing aggregated data
All above
huge delay
categorizing
small delay
IoT
no delay
avoiding central dataset through the use of p2p
decentralized „quasi"-p2p analysis
messages through telegrams
solution through area requests
trust for forwarding server
only near

List of Figures

List of Tables

Bibliography

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

[2] 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.

[3] 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.

[4] 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.

[5] T. Guardian. Accessed: 2019-04-10.

[6] 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.

[7] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. "Enhancing security and privacy in traffic-monitoring systems." In: *IEEE Pervasive Computing* 5.4 (2006), pp. 38–46.

[8] 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.

[9] W. A. Jabbar, M. Ismail, and R. Nordin. "Peer-to-peer communication on android-based mobile devices: Middleware and protocols." In: *2013 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO)*. IEEE. 2013, pp. 1–6.

[10] H. Kajino, H. Arai, and H. Kashima. "Preserving worker privacy in crowdsourcing." In: *Data Mining and Knowledge Discovery* 28.5-6 (2014), pp. 1314–1335.

[11] 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.

[12] 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.

[13] 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.

[14] M. Stolpe. "The internet of things: Opportunities and challenges for distributed data analysis." In: *ACM SIGKDD Explorations Newsletter* 18.1 (2016), pp. 15–34.

[15] 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.

[16] 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.

[17] 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.