On (The Lack Of) Location Privacy in Crowdsourcing Applications

Spyros Boukoros, *TU-Darmstadt;* **Mathias Humbert,** *Swiss Data Science Center* **(***ETH Zurich, EPFL***);** **Stefan Katzenbeisser,** *TU-Darmstadt, University of Passau;* **Carmela Troncoso,** *EPFL*

https://www.usenix.org/conference/usenixsecurity19/presentation/boukoros

This paper is included in the Proceedings of the 28th USENIX Security Symposium.

August 14–16, 2019 • Santa Clara, CA, USA

978-1-939133-06-9

Open access to the Proceedings of the 28th USENIX Security Symposium is sponsored by USENIX.

On (The Lack Of) Location Privacy in Crowdsourcing Applications

Spyros Boukoros[1], Mathias Humbert[2], Stefan Katzenbeisser[1,3], and Carmela Troncoso[4]

[1]*Department of Computer Science, TU-Darmstadt, Germany*
[2]*Swiss Data Science Center, ETH Zurich and EPFL, Switzerland*
[3]*Department of Computer Science and Mathematics, University of Passau, Germany*
[4]*SPRING Lab, EPFL, Switzerland*

Abstract

Crowdsourcing enables application developers to benefit from large and diverse datasets at a low cost. Specifically, mobile crowdsourcing (MCS) leverages users' devices as sensors to perform geo-located data collection. The collection of geo-located data raises serious privacy concerns for users. Yet, despite the large research body on location privacy-preserving mechanisms (LPPMs), MCS developers implement little to no protection for data collection or publication. To understand this mismatch, we study the performance of existing LPPMs on publicly available data from two mobile crowdsourcing projects. Our results show that well-established defenses are either not applicable or offer little protection in the MCS setting. Additionally, they have a much stronger impact on applications' utility than foreseen in the literature. This is because existing LPPMs, designed with location-based services (LBSs) in mind, are optimized for utility functions based on users' locations, while MCS utility functions depend on the values (e.g., measurements) associated with those locations. We finally outline possible research avenues to facilitate the development of new location privacy solutions that fit the needs of MCS so that the increasing number of such applications do not jeopardize their users' privacy.

1 Introduction

Crowdsourcing is a participative online activity in which the undertaking of a task is outsourced to a group of individuals [29]. This new paradigm of distributing a fragmented task, is an efficient, scalable business model that allows the cheap (or often free) massive collection of data. Indicative of the growth of this data collection methods is the appearance of over 2,000 crowdsourcing platforms [1, 56] in the last years [83]. Furthermore, according to recent industrial reports [40], in the last decade, 85% of top global brands have already adopted crowdsourcing, and in 2018, 75% of the world's highest performing enterprises would use crowdsourcing. For instance, Google [2], Microsoft [3] and Mozilla [4] use crowdsourcing to build WiFi location databases.

A driving force of the crowdsourcing ecosystem growth is the widespread adoption of smart mobile devices, which enable users to collect geo-located data on their devices and share it with central servers to attain a particular objective. Mobile crowdsourcing applications (MCS) have millions of users around the world. For instance, OpenStreetMaps [5], a map generation project from contributed GPS points, reports 4.3 million users in 2018,[1] with 1 million active map editors contributing over 4 billion GPS points. Similarly, OpenSignal [6], a popular network-measuring application, reports over 20 million users.[2] Safecast [7], a citizen science project collecting environmental data, currently reports over 75 million measurements from approximately three thousand users. Many other applications are available [4, 6, 8–18].

MCS can bring great benefits for organizations and society. However, the collection and sharing of geo-located data raises serious privacy concerns, as demonstrated by scandals related to the publication of data by fitness applications [19, 20] or irresponsible data analysis by transportation companies [21]. Location data can be used to identify points of interest (POIs) [49, 52, 64], infer users preferences, or de-anonymize anonymous traces [89]. This risk increases when considering auxiliary publicly available information [30, 63, 70], and persists even when protections are put in place [77, 78].

Over the last decade, the research community has proposed a vast number of LPPMs to address these issues [76], some of which can provide strong differentially private guarantees [28, 34, 48] and even offer optimal utility [33, 73]. Even though it seems like the location privacy question is technically solved, the reality is that these LPPMs *solely focus on one use case*. They are generally geared towards LBSs in which users sporadically reveal their location in return for a service (e.g., to find nearby restaurants). In this context, utility is user-centric and hinges on the precision of the reported locations. In MCS applications, on the contrary, geo-located data is often shared continuously and over long periods and, while

[1] https://wiki.openstreetmap.org/wiki/Stats
[2] https://opensignal.com/methodology#over-20-million-users-of-our-app

the data utility is still correlated with the location precision, it is foremost tied to the values of the measurements reported at these locations (e.g., WiFi signal strength, or radiation level). Moreover, MCS utility cannot be captured with a user-centric approach as, by definition, MCS benefits from aggregating data collected by a large amount of users.

In this paper, we conduct the first in-depth evaluation of the effectiveness of LPPMs in the context of MCS. We use two representative applications, Safecast [7] and Radiocells [9], which make their contributors' data publicly available on their websites and which have very different utility functions. We propose two new privacy metrics based on statistical measures developed for binary classification and information retrieval to capture the privacy gain provided by the LPPMs with respect to the identification of areas and points of interest. We also consider new utility measures that, instead of relying on distance-based errors, quantify the accuracy of the aggregate values of data collectively generated.

The results of our experimental evaluation on real data contradict common beliefs regarding the privacy-utility trade-off offered by different LPPMs. First, location hiding methodologies, which in LBSs help concealing trajectories [60], do not bring any privacy benefits to MCS users. This is mainly because, in MCS, the volume of geo-located data is larger and contains points reported over long periods of time (more than a day). Second, differentially private mechanisms [28] offer good protection only for very strong parameters, and even when they are optimized for utility [33], they dramatically perturb the radiation measurements. For instance, in Safecast, we observed that it tremendously changed some areas' radiation levels, urging people to evacuate a place, and completely hindered the ability to localize radiation hotspots (location with elevated radiation). Finally, generalization techniques, usually dismissed in LBSs because of their poor utility, offer one of the best privacy-utility balance in MCS.

In summary, existing LPPMs are not well aligned with the needs of MCS applications. Therefore, new research is needed to approach the design of optimal LPPMs based on collective, value-based, utility metrics instead of user-centric, location-based utility.

Our contributions can be summarized as follows:

✓ We propose novel privacy and utility metrics suitable to evaluate the performance of LPPMs for MCS data publishing patterns.

✓ Using real data collected from two representative MCS applications, we show that existing LPPMs impose too high utility price and that many of them do not even provide good privacy guarantees in the context of MCS.

✓ We discuss technical and non-technical countermeasures to improve the privacy protection of MCS users.

2 Mobile Crowdsourcing Applications

In this section, we introduce the two crowdsourcing applications studied in detail in this paper.

2.1 Safecast

Safecast [7] is a volunteer-centered organization whose goal is to monitor the global radiation levels and detect abnormalities in near real time. Safecast crowdsources the collection of radiation data by providing users with devices that collect radiation measurements every five seconds.

Safecast dataset. This dataset contains 64.2 million measurements from 608 users, collected from 2011 to 2017. Radiation measurements contain the user's name, a unique user ID, the device's ID, latitude and longitude, a UTC timestamp, and the radiation value and units. No registration is required to access these data and Safecast's privacy policy[3] states that to enable flexibility "Anyone is free to use with no licensing restrictions". For our experiments we removed IDs corresponding to organizations, malformed entries, and converted all UTC times to local. After this process, the dataset has almost 56.7 million measurements from 540 users.

Safecast utility. The Safecast project uses the collected data to study different phenomena related to radiation. In this paper, we consider two of the main uses of the data.

First, we consider the interactive map to visualize radiation published on Safecast website. Safecast computes the visualized radiation levels from the crowdsourced measurements as follows. For a given region of interest, Safecast filters the measurements within the region and computes the average radiation at each location over the last 270 days. Second, they discretize the area to 2.25 million grid points (1500 discrete locations per axis). They create the displayed map using nearest-neighbor interpolation on the averaged radiation measurements associated to the points of the grid. The reported radiation is measured in counts per minute (cpm), expressing how many ionized particles are detected per minute by a monitoring instrument. This use case, which relies on averaging and interpolation, represents a setting in principle amenable to noise in the data.

Second, we consider the detection of *hotspots* – specific areas where radiation is above a pre-defined threshold. These hotspots indicate locations where radiation could be harmful for public safety. Once identified, Safecast might send experts to perform on-site examination to better understand the causes and consequences of such dangerous zones. Therefore, it is crucial that the localization of hotspots is accurate.

2.2 Radiocells

Radiocells [9] is a community project whose goal is to provide an open-source alternative to commercial, closed source, geo-

[3] https://blog.safecast.org/faq/licenses/

location databases for cell towers and wifi base stations. They also aim to provide raw telecommunication infrastructure data for use in diverse scientific studies. Radiocells crowdsources the collection of measurements via a mobile application called 'Radiobeacon'.[4] With this application, users continuously collect measurements as they perform daily activities. Users choose when to start and stop measuring, and when to upload the measurements to the Radiocells server. Furthermore, they can select a specific area where measurements will not be recorded, e.g., to protect their home locations. We do not study the impact of this defense in this paper, but previous work shows that it is rather fragile [57].

Radiocells dataset. The raw data uploaded to the server is publicly available for download. It is licensed under Creative Commons Attribution-ShareAlike 3.0 Unported and ODbL licences aimed at not restricting the use of the data.[5] Amongst other information, the measurements include: signal strength, cell (antenna) ID, location, timestamp, and smartphone model, software, OS version, and manufacturer. In an effort to preserve users' privacy, this dataset does not contain usernames. However, the combination of the smartphone characteristics, the location, and the network provider is likely to represent a quasi-identifier. We downloaded data for 2013 to 2017, obtaining 25 million measurements. To separate users' measurements, we grouped the measurements according to phone manufacturer, phone model, country and network operator. We obtained 998 potential unique users, of which we only kept those that had more than 100 measurements. We also removed users with spatial inconsistencies, i.e., we removed all users whose speed between two contiguous measurements was greater than 200 km/h. The dataset finally contains 568 users and about 4 million measurements.

Radiocells utility. Amongst other purposes, the Radiocells data can be used to geolocate antennas. Such information is useful to enable scientific studies about antennas distribution and signal quality in specific places. Contrary to Safecast, Radiocells does not provide documentation, nor provide code indicating how they produce their map of antennas. Thus, we use the location function described by OpenCellID [8], another crowdsourcing project with the same goal, which defines the location of an antenna as the average of the latitudes and longitudes of the measurements referring to this antenna.

3 Protecting Location Privacy in MCS

In this section, we describe the existing LPPMs we evaluate in our study. These LPPMs are designed for LBSs settings, which are different than MCS in two aspects. First, LBSs aim at fulfilling an individual need related to one user's location (e.g., find nearby restaurants), while MCS aims at fulfilling a common objective through collaborative measurements. Second, LBSs can often work with sparser geo-located data (just few points per geo-located query) than MCS, which requires continuous data collection and in a larger volume.

3.1 Defenses

We consider three type of LPPMs [65, 81]: (i) spatial obfuscation, (ii) hiding, and (iii) generalization. We do not consider the use of dummy locations or synthetic data [32, 35]. Both approaches focus on producing plausible artificial locations, but to the best of our knowledge there is no proposal that provides the means to generate measurements (or other values) to be associated to these locations. In fact, we argue that generating fake measurements, even using prior information, is bound to pollute the real-time measurements that these applications aim at collecting.

Spatial obfuscation. The state of the art in spatial obfuscation, which perturbs reported locations with noise, is *geo-indistinguishability (GeoInd)* [28]. This mechanism adapts differential privacy to location data, providing privacy guarantees independent from the adversary's prior information. This approach is widely used in the literature [22, 48, 62, 68, 75, 88]. Following the original definition in [28], we obfuscate locations by adding planar Laplacian noise. The magnitude of this noise is controlled by the parameter $\varepsilon = l/r$ which guarantees that the ratio between the probabilities of two points being the real location in an area of radius r is at most l.

Release-GeoInd. As with any differentially private mechanisms, in GeoInd the level of privacy decreases linearly with the number of reported locations. To address this limitation we implement a mechanism inspired by the predictive approach proposed in [34]. This defense reports a new noisy location if, and only if, the user has moved at least z meters away from his previous location. Otherwise, it repeats the last reported location. We call this approach "Release-GeoInd".

GeoInd-OR. Remapping[6] obfuscated locations to popular places according to prior knowledge on users' movements can offer optimal utility without reducing privacy [33, 73]. We complement GeoInd with the remapping approach in [33]. We refer to this approach as "GeoInd-OR".

Hiding. This defense achieves privacy by suppressing some of the users' locations [60, 61]. The released locations are *not* perturbed. We consider two hiding strategies: (i) a "Random" strategy in which users release a random subset of their points, and (ii) a "Release" strategy in which users only reveal a new point when they have traveled at least x meters away from the previously reported location.

Generalization. This defense reduces the precision with which locations are reported [31, 55]. We implement this approach by reducing the precision of the reported GPS coordinates [65]. We denote this defense as "Rounding".

[4] https://f-droid.org/packages/org.openbmap/

[5] https://radiocells.org/license

[6] A remapping g is a function $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ that maps an output $z \in \mathbb{R}^2$ to another output $z' \in \mathbb{R}^2$ according to the probability density function $g(z'|z)$.

3.2 Measuring Privacy

Location privacy metrics in the literature are mostly based on a function of the distance between the real location of the user and the one inferred by the adversary [80, 81]. This function could measure the *correctness* of the adversary's inference (e.g. using, Hamming or Euclidean distances [81]), or the *uncertainty* of the adversary regarding the user's location (e.g., using entropy [73]). These metrics are very well suited for the case of LBSs, where users release one location per query, and the adversary tries to infer that location. However, they are hard to use in the MCS setting, where the adversary has access to locations released continuously over several days. In this case it is hard to establish between which points to compute a distance, or across which points to compute probability distributions for entropy-based metrics.

We also argue that the metrics above do not capture privacy in a manner understandable by users and developers of crowdsourcing applications. How much privacy is an error of 10 meters or 500 meters? It is clear that one is larger than the other, but not how much privacy they provide regarding the potential inference of sensitive information. Even more complicated is the case of entropy, whose units of measurement – bits, nats, or hartleys – are rarely known, let alone interpretable, by layman people.

Privacy gain. We propose to quantify privacy as the loss of adversarial inference power regarding two privacy dimensions understandable by users: geographical area and POIs. To quantify this loss, we use two well-established statistical measures: precision and recall. The former captures the increase in privacy when, after a defense, the adversary identifies many false candidate locations along with the user's real whereabouts. Here, the adversary has low *precision* ($\frac{TP}{TP+FP}$, where TP and FP refer to *true positives* and *false positives*, respectively). The latter captures the increase when, after the defense, the adversary cannot correctly identify the original locations visited by the user. Here, the adversary has low *recall* ($\frac{TP}{TP+FN}$, where FN refers to *false negatives*).

Spatial privacy gain. Spatial privacy considers the geographical *area* in which the adversary infers the user can be. We define the true positives (TP) as the intersection of the areas where the user can be before and after applying the defense (i.e., the area inferred by the adversary that corresponds to the user's real location). Similarly, we define the false positives (FP) to be the set difference of the area after the defense and the area before the defense (i.e., the area inferred by the adversary where the user was not present), and false negatives (FN) as the set difference of the area before the defense and the area inferred after the defense (i.e., the area where the user has been but that is missed by the adversary).

POI privacy gain. In reality though, the geographic area itself may not reflect users' privacy [80]: if there is only one point of interest in a large area, privacy should be low; and in small areas with many POIs (e.g., a block in a city), privacy should be large. We propose a complementary metric based on POIs. In this case, true positives (TP) are the POIs in the intersection of areas before and after the defense is applied. Similarly, false positives (FP) are POIs identified after the defense that were not present before, and false negatives (FN) are the POIs inferred initially that are missed after the defense.

Table 1: Safecast (top) and Radiocells (bottom) measurements per region. Vulnerable users are those with at least one cluster.

Region	Users	Measurements	Average per user	Standard deviation	Vulnerable users
Tokyo	30	2,701,367	90,046	203,576	24 (80%)
Fukushima	104	7,765,773	74,671	260,671	65 (62%)
World	540	56,655,768	105,504	70,954	349 (65%)
World	568	3,710,547	6,532	17,312	91 (16%)

3.3 Measuring Utility

Similarly to privacy, in LBSs, utility is measured as a function of distance between real and obfuscated locations of one user. This is unsuitable for MCS where location depends on the precision of the aggregate of multiple users' geolocated measurements. We now introduce the utility metrics used in our evaluation.

Distance-based. We call distance-based metrics those associated to LPPMs in the context of LBSs. In our experiments, we use the per-location haversine distance[7] between original and obfuscated locations.

Aggregate statistics. Most MCS providers are interested in aggregate statistics computed over individuals' contributions. This is the case for Safecast and Radiocells, where the radiation map, respectively the coordinates of the antennas, are derived from average measurements of MCS users. In our evaluation, we consider as MCS utility metrics the actual utility functions of the projects as described in Section 2.

4 Existing LPPMs Performance in MCS

4.1 Experimental setup

We experiment on all data available from Safecast and Radiocells. For Safecast, we additionally consider two regions in Japan with very different radiation profiles: Tokyo, where the radiation profile is quite uniform, and Fukushima, where the nuclear incident at the Daiichi power plant [23] in 2011 created areas with elevated radiation. Table 1 summarizes the statistics (number of users, total amount of measurements, and measurements per user) of the regions under study.

We evaluate the privacy gain and the utility loss of an LPPM as follows:

[7] Distance between two points on a sphere given their longitudes and latitudes.

Step 1. Adversary's inference. Inspired by previous works, we use clustering to implement inference on the regions and the points of interest for all users. [36, 42, 46, 49, 59, 64, 69, 86]. Concretely, we use the density-based clustering algorithm (DBSCAN) [46]. Contrary to other clustering algorithms (such as K-Means), DBSCAN is robust to noise and outliers and does not require to specify the number of clusters a priori (see Appendix A.1). We keep the five clusters with the highest number of points, and *we consider their total area as the geographical area input to the Spatial Gain metric*. Once clusters are identified, we use the OSM API[8] to find the POIs in the clusters of the targeted user. *We consider all points in the top five clusters as input for the POI Gain metric*. Table 1 reports the percentage of users *vulnerable* to our attacks *before* the defenses are applied, i.e., the percentage of users for which we find at least one cluster. For Safecast-Tokyo, we only report statistics for the 30 users considered when using GeoInd-OR (see Section 4.3).

Step 2. LPPM Application. We apply the LPPM to all users' data and repeat the actions in Step 1 to infer their regions and points of interest. Note that when Rounding to 2 or 3 decimals, obfuscated locations are separated by approximately 1,100 meters and 110 meters, respectively. Thus, our parametrization of DBSCAN is bound to not find any clusters. However, an adversary would know that given an obfuscated point, the actual location of the user is within a square of size 110, resp. 1,100 meters, centered in the reported location. Thus, for this case, instead of using DBSCAN clusters, we pick the squares of the respective sizes around the five most frequently reported obfuscated locations.

Step 3. Privacy gain. We compare the area (in square kilometers) of the clusters before and after the LPPM to compute the Spatial privacy gain, and the POIs inside the clusters for the POI privacy gain.

Step 4. Utility loss. In the case of aggregate statistics, the utility loss is application dependent. For Safecast, we consider the absolute difference in cpm per grid point between the radiation values on the application's interactive map (see Section 2.1), before and after the LPPM. In Radiocells, we consider as utility loss the distance between the location of the antennas before and after the LPPM.

4.2 Validating the Inference Strategy

We now validate the suitability of DBSCAN as strategy for inferring areas and points of interest in the context of MCS. Specifically, we test its suitability to identify workplaces on data from Safecast and OpenStreetMaps. As both projects' public data contain identifiable information about their users, we can validate the inferences against information available on other online platforms. We choose workplaces for ease of validation, but we note that it is just one of many inferences that could be done using location data [42]. Our results below confirm that DBSCAN is a suitable choice as basis to compute areas and POIs to input in our privacy metrics.

[8] https://wiki.openstreetmap.org/wiki/Overpass_API

Table 2: Safecast dataset statistics.

Measurements	Users	Avg measurements	Avg days
<10k	213	3,331	5
10k-100k	230	38,341	20
100k-1M	87	270,387	105
>1M	10	1,958,760	632

Ethical considerations. For these experiments we do not collect any personal data other than that made publicly available by the MCS projects. We have limited our inferences to the minimum to validate the suitability of DBSCAN. We only report aggregated or anonymized data such that no individual's data is exposed. We have notified the service providers about our findings, and we have shared our code with them so that they can make informed decisions regarding improvements of the privacy situation. Our code is open-source so that it can also be used by other crowdsourcing applications and improved by the research community [24]. This procedure has been approved by EPFL's Human Research Ethics Committee (HREC).

Safecast. To evaluate the effectiveness of DBSCAN in different situations, we split the users in the dataset into four groups according to their amount of measurements they report. For each group, Table 2 shows the number of users, their average amount of measurements, and the average number of days in which they took at least one measurement. From each group we select as targets for inference the 10 users with the most measurements that provide their real names. Since in the group with >1M there are only 4 users with real names, we end up with 34 target users in total. This allows us to manually validate our inferences in reasonable time.

Identifying workplaces. We run DBSCAN on every users' measurements during working hours (Monday to Friday from 9AM to 5PM). We configure DBSCAN to find clusters with at least 80 points separated by 60 meters and, if no clusters are found, we increase the distance by 30 meters (up to 120 meters maximum) and decrease the number of points by 15 (down to 35 points). These parameters have been chosen empirically to optimize the adversary's success, see Appendix A.5. To keep the manual analysis feasible, we only consider the five clusters with the highest number of points.

We expect that the users' workplace is one of the POIs within the inferred geographic area. In many cases, however, this area is large and contains many POIs. To ease manual validation, we use X-means clustering [74] to split these large clusters, and consider as POIs the centroids of the two largest subclusters. We end up with at most 10 POIs per user. We use the MapQuest API [25] to obtain these locations' addresses and, if existing, the names of the businesses at those coordinates. We recall that in our LPPM evaluation below,

we consider all points in the clusters as input for the POI Gain metric. This represents a resourceful adversary that can afford checking manually all the points and filter out those corresponding to businesses. We note that considering more points could cause more false positives, but the semantics of locations often makes it easy to filter these out, e.g., lakes or parks can be usually discarded as workplaces.

Once we have candidate workplaces, we validate them using social networks such as Twitter or LinkedIn, or the users' personal webpages. We note that 9 of our target users did not have a publicly available profile or had too common names to find their correct information, thus we could not validate their inferred workplaces. Overall, we recover the workplace of 35% of the target users. This result is consistent across the groups: 40% of the users with less than 10k measurements, 20% of the users with 10k-100k measurements, 50% of the users with 100k-1M measurements, and 25% of the users with more than 1M measurements. We conclude that DBSCAN performs well for POI identification irrespectively of the amount of data shared by the users. Surprisingly, this means that privacy seems not correlated to the volume of data made available to the adversary. On the contrary, it seems to be highly dependent on the collection patterns of the users. We observe that people fall in one of two categories: (i) Those who travel to different places with the goal of obtaining measurements, whose work addresses cannot be inferred; and (ii) those who measure radiation during their daily activities, whose work place we can find. The Safecast co-founders, who are the top contributors in terms of data points, fall in the first category, explaining the lower inference power for users with more than 1M measurements.

Our results confirm recent findings in the literature regarding personal information inferences from location data [42, 45]. Yet, we want to stress that the threat may be worse for MCS, due to the volume of data exposed by participants. For reference, Safecast's lowest contributing group has on average 3k measurements per user (see Table 2) while in the Twitter analysis performed by Drakonakis et al. [42] only the top contributing users (less than 0.06%) have more than 3k geolocated tweets. Thus, even if the number of MCS users is not as large as social networks' users, we expect a significant fraction of them to be vulnerable to privacy attacks.

Other POIs. A deeper analysis of the times and semantics of the POIs identified by DBSCAN revealed further information about Safecast users. Among others, we could infer two users' membership to specific organizations: one member of the Scientology church who reported many points from the Church of Scientology Celebrity Centre in a major city; and a Masonic lodge member who regularly visited the lodge headquarters. We could verify this information online for both users. We also identified two work-related activities: a US-based scientist working on a project about radiation around a lake in the Southern part of the US, and a photographer working in a Japanese city. We validated these inferences using Research Gate and the webpage of the artist, respectively. Finally, we could follow the education steps of a European PhD student. Her points of interest over time reveal the university where she obtained her master's degree, an exchange with another European university, and the university where she is completing her doctoral studies. We verified these facts on her CV available online.

OpenStreetMaps. Contrary to Safecast, OSM does not have an open API for accessing users' data. Yet, traces from users who have chosen to make their data available can be easily obtained from OSM's website.[9] To minimize the impact on OSM servers, and comply with their non-crawling policy, we manually downloaded data for 30 users with a large amount of contributions,[10] of which 17 used their actual names (or indicative nicknames). Although the majority of the points in the dataset were rather old (most of them at least 7-8 years old), we were able to verify previous workplaces for 3 of the 17 users (17%). We note that, for some users, we found out that they did not have a standard place of employment during data collection period (e.g., students). However, for *all* users, their POIs were within the area where they worked or lived. We used this fact to infer two of the users' short vacation trips which we manually verified with information publicly accessible from their social media accounts.

4.3 Privacy Gain

Defenses implementation. For the GeoInd defense, we set the privacy parameter $l = \ln(1.6)$, and use radius $r \in \{50, 150, 300\}$ meters which yields $\varepsilon \in \{0.01, 0.003, 0.001\}$. Remapping the locations for the LPPM GeoInd-OR requires computing the posterior probability for every candidate location. This operation is rather costly when the number of locations being considered grows. To keep a reasonable experimentation time, we only test GeoInd-OR for the Tokyo region in the Safecast dataset. We use 80% of the users to construct the prior probability distribution describing users' movements, and the remaining 20% to evaluate the effectiveness of the approach. We chose this 20% manually to keep a balanced testing set. It is composed of the top 10 users with many (more than 50k), moderate (between 10k and 50k), and few (less than 10k) measurements. Finally, for the Release-GeoInd mechanism, we use $l = \ln(1.6)$, $r = 50$ meters, and we select the distance between released locations to be $z \in \{30, 60, 90\}$ meters. We provide details about the implementation of these LPPMs in Appendices A.2 and A.3.

We implement the Random mechanism tossing a biased coin every time a location is about to be reported. The bias is set so that users release on average 40%, 60% or 80% of their measurements. For the Release mechanism, we sort all the locations reported by a user in chronogical order, and

[9] https://www.openstreetmap.org/traces
[10] http://resultmaps.neis-one.org/oooc

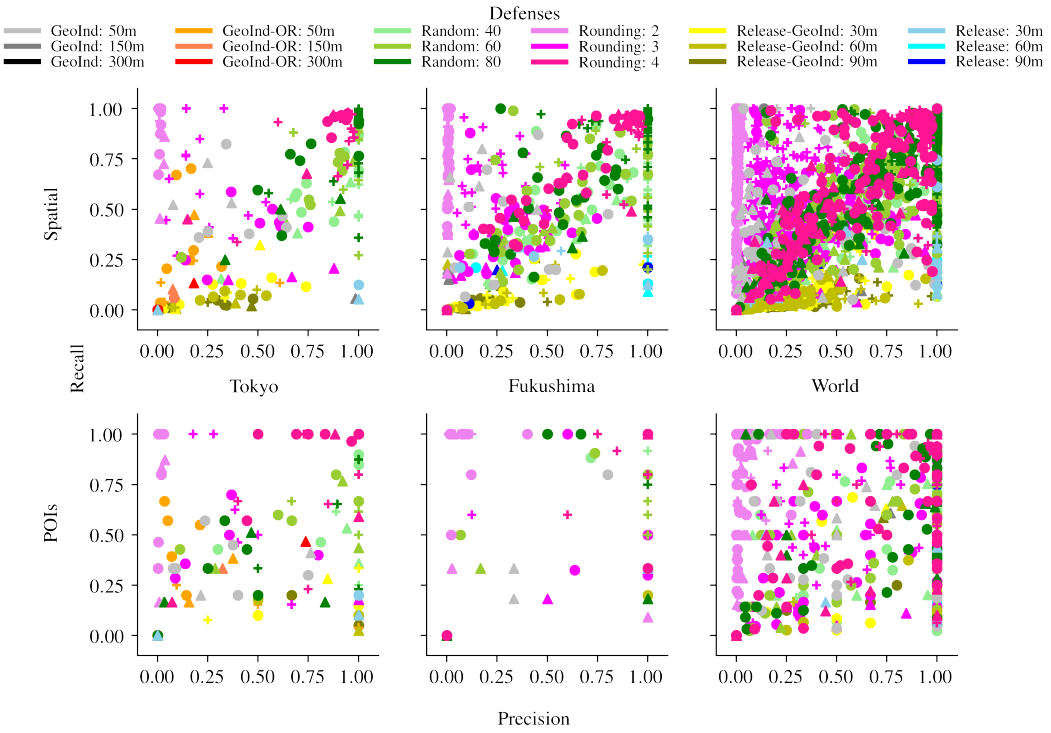

Figure 1: Safecast privacy gain: Spatial (top) and POIs (bottom). Amount of measurements per user + : <10k, ● : [10k,50k], ▲ : >50k. Each point on the graphs represents one user.

release a new location only if it is separated by (at least) $x \in \{30, 60, 90\}$ meters from the previously reported one. If two locations are less than x meters apart but in different days, we release them both.

Last, we implement Rounding by rounding to 2, 3, or 4 decimals the latitude and longitude of the users' locations. Effectively, this reduces the location accuracy to roughly 1,100 meters, 110 meters and 11 meters, respectively.

4.3.1 Safecast

We first evaluate the privacy gain of the LPPMs in the Safecast dataset. Figure 1 shows the Spatial (top) and POI (bottom) gain for Tokyo, Fukushima, and the whole world. (Figure 12 in the appendix shows the results for each of the defenses separately for the whole world.) The x-axis represents precision, and the y-axis recall. Each point in the graph represents a user, and the markers' shape indicate the amount of measurements she contributes. The colors represent the LPPMs. To compute these graphs, we configure DBSCAN to find clusters with at least 75 points separated by at most 30 meters (roughly the size of a small building). As in [42], we also require that, for each cluster, users either stay more than 30 minutes, or visit it for at least two days. A first reason to fix these parameters is to evaluate the gain for all users under the same conditions. A second reason is that the loose parameters used in Section 4.2 can yield very sparse clusters with few points that are hard to break by removing or perturbing locations. Thus, the defenses would perform equally bad and we would gain little information about their properties. Tightening the parameters reduces the work inference success to 21% (some clusters are not found), which still represents a significant risk.

Defenses that provide large gains result in points close to the figure axes. Points near the y-axis indicate low precision, i.e., cases in which the adversary correctly identifies some (or even all) of the true locations but also inferred many other wrong locations. Points near the x-axis indicate low recall, i.e., cases in which the adversary correctly identifies some real locations, but misses many others. Unsurprisingly, we observe a high variance in the defenses' performance since it is highly dependent on the user behavior. However, it is possible to identify some trends.

We first discuss the Spatial privacy gain (Figure 1, top). For the least privacy-preserving parameter ($r = 50$m), GeoInd significantly decreases the number of vulnerable users (grey points in the figure) from the values reported in Table 1. The reduction is 50% for Tokyo (from 24 vulnerable users to 12), 45% for Fukushima, and 45% for the whole world. When the mechanism is strengthened ($r = 300$m), GeoInd adds so much noise (see Figure 10 in Appendix A.4 for reference) that no users are vulnerable after the defense. In summary, GeoInd seems to provide fairly good privacy gain in Tokyo

and Fukushima. Yet, when we look at the whole dataset, it becomes clear that the protection provided by GeoInd is highly dependent on the users' movement patterns.

The Release-GeoInd (yellow) mechanism works generally better than GeoInd. Even though more users are vulnerable (only between 4% and 13% of the users become not-vulnerable) and the adversary obtains reasonable precision, it yields very low recall. This is because in this method users keep reporting the same obfuscated location until they move. This repetition results in clusters being found on fake locations that often do not overlap with the original ones. This reduction becomes more significant as the defense is configured to provide more privacy (larger z).

GeoInd-OR performs slightly better than vanilla GeoInd. This is because the remapping results in points being repeatedly mapped to popular places causing the generation of clusters around those not-real locations.

Similar to vanilla GeoInd, the Release mechanism (blue) significantly reduces the number of vulnerable users – by more than 50% even for the least conservative parameter. However, when precision is very high, i.e., when a cluster is found, it corresponds to a real location. The reason is that even though the user hides many points, if a location is visited regularly, the user will eventually report enough points around this location to make the cluster identifiable by the adversary.

The Random hiding mechanism (green) does not perform well. First, it reduces the number of vulnerable users less than other defenses (10% decrease in Tokyo, 27% in Fukushima, and only 5% when considering the whole world). From the vulnerable users only a handful obtain good protection. We could not find a clear pattern to predict which movement profiles would best benefit from this defense. For many users, especially those with a few points, removing points at random still yields high precision as the few measurements are very localized. Overall, we do not notice much influence of the fraction of hidden points on the privacy of the users.

Finally, the protection provided by Rounding (pink) depends on the rounding parameter. Keeping 4 decimals reduces accuracy by just 11 meters. Therefore, the adversary finds roughly the same clusters, i.e., for many users we observe high recall and precision after the defense (especially in Tokyo and Fukushima). On the contrary, rounding to 2 or 3 decimals significantly increases the size of inferred spatial areas, which leads to variable recall (depending on the users' movement patterns) and low precision.

Regarding the POI privacy gain (Figure 1, bottom), a first observation is that the amount of users vulnerable to the attack, i.e., points in the graph, is lower. This is because for many users the identified clusters do not contain any POI (according to the OSM API). Second, for the users who have POIs in their clusters, both recall and precision are higher than in the Spatial gain. This is because many of the large clusters that contribute to the low Spatial precision do not have POIs and thus do not contribute to the confusion of the adversary when identifying particular POIs. Furthermore, the clusters that the adversary finds after the LPPMs may cover a smaller area than the original clusters, but still contain most of the users' initial POIs. This provides a higher POI recall than Spatial recall. Third, in this case we observe a significant difference between Tokyo and Fukushima. The reason is twofold. First, the Fukushima prefecture is much larger than the area of Tokyo we consider. Second, Fukushima is a rural area and thus contains fewer POIs than Tokyo where even small clusters have many places of interest.

These observations reinforce previous insights that solely considering the spatial dimension may provide a false perception of privacy [80]. Considering a POI-privacy measure is necessary for providing a comprehensive picture of the privacy threat users face in MCS applications. We note that this perception also depends on DBSCAN parameters, which define the size of the regions found, and consequently the number of POIs, increasing the manual effort of the adversary. We discuss this effect in Appendix A.5.

Impact of the amount of measurements on privacy. We present in Figure 2 the Spatial gain for the three best LPPMs (all parameters combined) split by the amount of measurements users contributed. We discard Rounding 4 as it does not provide any privacy. We see that all LPPMs provide low precision and recall regardless of the users' contribution volume. The exception is Rounding which, as explained above, by definition provides variable recall and low precision.

Counterintuitively, the LPPMs perform worse for users who contribute fewer points. This is because the attack constructs more, and larger (on average 10 times bigger), clusters for people who share many points than for those sharing fewer points. These clusters are split after the LPPMs are put in place, as some reported locations are moved away from their original clusters while other measurements, perturbed with noise, concentrate to new places forming wrong clusters. For Rounding, where every cluster created after the LPPM has roughly the same size, users with a few measurements have higher recall because their initial small clusters are often covered by the large regions resulting from the LPPM.

Thwarting workplace inference. Finally, we evaluate the effectiveness of the different LPPMs at hiding workplaces. Recall that, without protection, we can identify the workplace of 21% (7 out of 34) of the users. Five defenses, GeoInd, Release-GeoInd, Release, and Rounding 2, protect all users from inferences. Random hiding requires heavy sampling to be effective (hiding only 20% permits the identification of 6 workplaces, and hiding 40% still reveals 1). Finally, unsurprisingly, Rounding to 4 decimals does not protect against work inference, and Rounding with 3 decimals only hides one workplace out of 7.

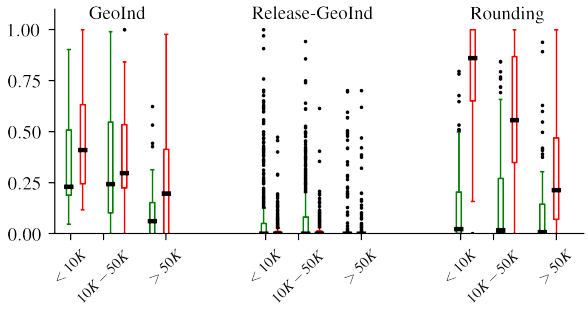

Figure 2: Precision (green) and recall (red), depending on the amount of measurements x per user for three selected defenses (all parameters combined).

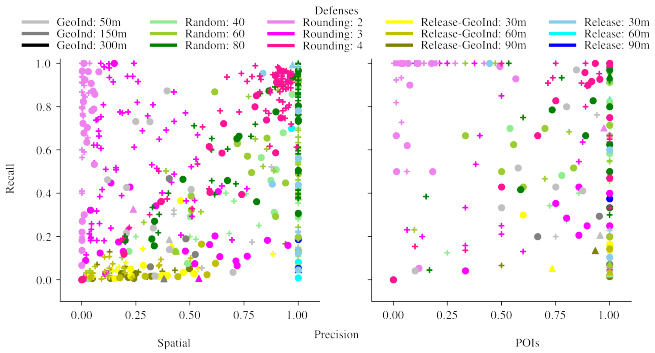

Figure 3: Spatial privacy gain (left part) and POI privacy gain (right part) in Radiocells. Amount of measurements per user + : <10k, ● : [10k,50k], ▲ : >50k. Each point on the graphs represents one user.

4.3.2 Radiocells

Users in Radiocells have on average fewer measurements than those in Safecast, and clustering requiring 75 points yields very few clusters. Hence, for this dataset we loosened the DBSCAN requirement to 25 points per cluster.

We see in Figure 3 that GeoInd-based mechanisms behave similarly to the Safecast case in terms of Spatial gain: GeoInd provides highly variable protection, and Release-GeoInd yields low recall while precision depends on the user behavior. Vanilla GeoInd decreases the number of vulnerable users by 14%, and Release-GeoInd by 2%. Given that only 16% of the users were initially vulnerable, this reduction is significant. For the hiding mechanisms, the Random and the Release mechanisms decrease the number of vulnerable users by 7% and 14%, respectively. For the vulnerable users, contrary to Safecast, these mechanisms consistently yield high precision, i.e., they offer poor privacy protection for Radiocell's users movement profiles. Finally, the Rounding mechanisms with parameters 2 and 3 offer reasonable privacy. Regarding POIs, we observe similar behavior to the Safecast dataset.

Overall, the results in Radiocells are consistent with our findings in the Safecast dataset, confirming the trends regarding the LPPMs behavior in the MCS setting.

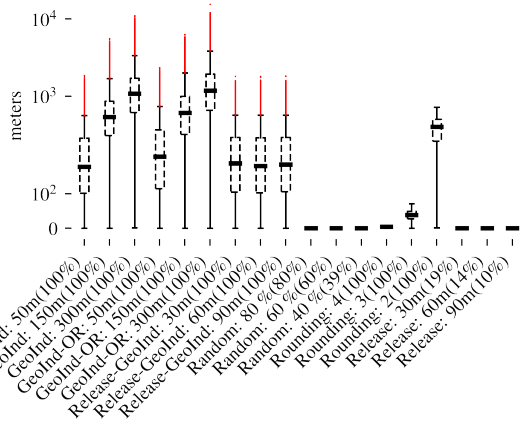

Figure 4: Measurement error in Tokyo using a distance-based metric. This can be interpreted either as privacy gain or utility loss.

4.4 Privacy-Utility Trade-Off

4.4.1 Safecast

Distance-based metric vs Aggregate statistics for MCS. We first evaluate the utility loss incurred by the LPPMs measured using the LBS-oriented distance-based metric described in Section 3.3. This utility metric is based on the distance between reported and real locations, but disregards the (radiation) values that Safecast cares about. Figure 4 displays the results for users in the Safecast-Tokyo dataset. The y-axis indicates the distance in meters, and the x-axis the LPPM and the percentage of points that are released. Random and Release LPPMs, which add no noise, are the best in terms of error; and GeoInd LPPMs offer the worst performance as they tend to spread locations — sometimes more than a kilometer away from the initial measurements (see Figure 10 in Appendix A.4).

Next we consider the utility loss for aggregated statistics, i.e., utility measured as the difference between radiation values to be plotted on the generated map. We plot per grid-point utility loss for Tokyo and Fukushima in Figures 5 and 6, respectively. We observe that the loss is similar in both regions, though in Fukushima the median loss is slightly higher and there are more, and larger (up to 10^4 radiation offset with respect to the original value), outliers than Tokyo. Because of the interpolation step, in this case all GeoInd variants offer roughly the same utility loss on average. Still, Hiding and Rounding strategies offer better performance, with small

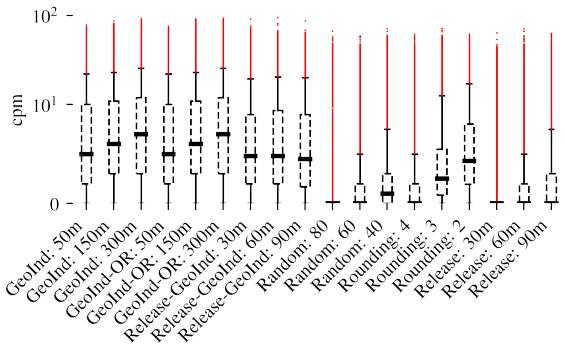

Figure 5: Absolute difference in Tokyo's radiation values with Safecast dataset.

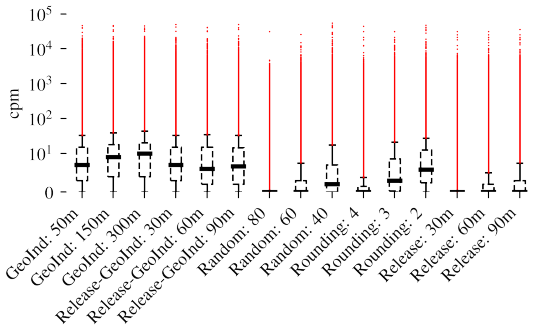

Figure 6: Absolute difference in Fukushima's radiation values with the Safecast dataset.

median error for the least protective parameters.

If we compare the distance-based results (Figure 4) to the aggregated statistics utility loss (Figure 5), we observe significant differences. First, the interpolation step results in LPPMs based on GeoInd to fare much better in terms of aggregates than in terms of distance. Second, distance-based metrics underestimate the utility loss of hiding LPPMs (Random and Release). While it is true that the released points have no error in distance, hiding points comes at a cost not reflected in the metric. This is made evident by the aggregated metrics, which show that the more points are hidden, the larger is the utility loss. We note that relying on Markov mobility models such as in [51, 81] could help interpolate the hidden locations. Yet, this would not help recover the (radiation) values attached to them and the utility loss would remain. For the generalization mechanisms, distance-based metrics consistently report larger median loss, but have less variance and less outliers.

In summary, distance-based metrics provide a very different perception of the LPPM performance than considering utility functions computed on the geo-located values, overestimating the performance of some methods (e.g., hiding strategies) and underestimating others (e.g., GeoInd-based LPPMs). We conclude that traditional LBS-oriented metrics are inadequate for measuring utility in MCS scenarios.

Semantic interpretation. The absolute difference in cpm of measurements before and after the defense gives a rough idea about the utility loss, but it is difficult to interpret. Is it significant? What is the effect of outliers? Does reporting the values after the defense have any implication on the danger for human health? To answer these questions, we study how the variance introduced by the defenses can change the interpretation of the risk at a given location. To this end, we rely on the cpm safety scale [26] provided with one of the top-seller Geiger counters (radiation measurement devices) on the market. This scale contains five categories:

- Category 1: 0-50 cpm. Normal radiation background.
- Category 2: 51-99 cpm. Medium level.
- Category 3: >100 cpm. High level.
- Category 4: >1000 cpm. Very high level, leave area.
- Category 5: >2000 cpm. Extremely high level, immediate evacuation.

Table 3: Danger category changes after applying Geo-Ind ($r = 300$ meters) in Fukushima.

Geo-Ind: 300m	**1**	**2**	**3**	**4**	**5**	**Number**
Original						**of points**
1	79.7%	19.3%	1%	0.003%	0.001%	1,354,110
2	41.5%	49.5%	9%	0.023%	0.01%	650,486
3	8.7%	35.9%	52.2%	2.3%	0.9%	229,848
4	2.5%	3.3%	49.3%	29.8%	15.1%	10,489
5	3.9%	1.7%	34.7%	29.3%	30.4%	5,067

We select the prefecture of Fukushima and two defenses that produce a good level of privacy: GeoInd 300m and Rounding 2. For each of the 2.25 million grid-points on Safecast's radiation map for Fukushima, we compute their radiation category according to the safety scale before and after each defense. For GeoInd 300m, which is of probabilistic nature, we repeat the procedure 10 times and report the average. We present the results in Tables 3 and 4. We observe that the majority of the points either stay in their original category or move to a nearby. However, we observe some extreme category jumps from the first category (safe radiation levels) to the fourth and fifth (high danger). For instance, GeoInd causes 53 places to be marked as dangerous instead of safe. Even more alarming, 283 locations that should be marked as extremely dangerous are marked as safe or slightly elevated (categories 1 and 2). On the contrary, the Rounding mechanism limits the number of extreme changes. For instance, there is a category jump from 5 to 1 and 2 only for 45 grid-points.

Why optimal remapping does not work for MCS. Even though GeoInd-OR was designed to increase utility while preserving privacy, we observe that, in the MCS case, utility roughly stays the same (Figure 5), and privacy slightly increases, both in decreasing the number of vulnerable users and in increasing the spatial gain. The reason for this mismatch is that this mechanism was designed in the context of

Table 4: Danger category changes after applying the Rounding mechanism (2 decimals) in Fukushima.

Rounding: 2 / Original	1	2	3	4	5	Number of points
1	89.3%	10.3%	0.3%	-	0.001%	1,354,110
2	30.2%	64%	5.8%	0.003%	-	650,486
3	0.7%	22.6%	74.8%	1.6%	0.3%	229,847
4	0.2%	0.01%	43.3%	39.6%	16.9%	10,490
5	0.9%	-	9.3%	42.1%	47.6%	5,067

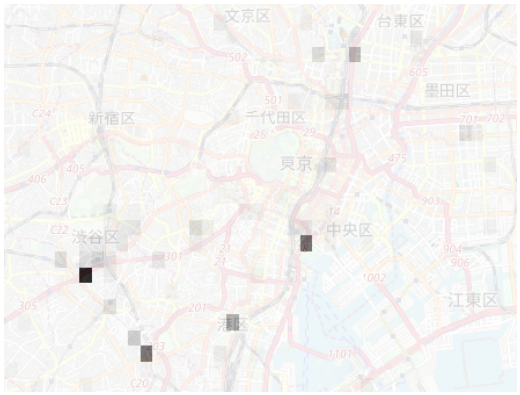

Figure 7: Prior probability of visiting locations in Tokyo (white - low probability, black - high probability).

LBSs, where remapping locations to places where the user is likely to be is bound to provide good utility on average. However, in Safecast, the utility does not depend on the locations themselves, but on the associated measurements. Remapping the location, however, concentrates measurements in these popular locations, effectively polluting the measurements. We illustrate this effect in Figure 7, which represents the prior probability of users' locations over all locations in Tokyo (low in white, high in black). In the low probability areas, most locations have the same probability, thus remapping has a randomizing effect. However, when there is a location with high probability, all locations are remapped to this popular location. We note that, while significantly hurting utility, this effect creates artificial clusters that reduce the adversary's precision and recall, thus increasing privacy.

The case of high precision measurements. Safecast also uses the crowdsourced measurements to monitor radiation *hotspots* that could be dangerous for public health. For this case, location precision is highly important, both to understand the dangers it can cause and to keep low costs if experts have to be sent to study the origin of the abnormality.

We study the impact of LPPMs on hotspot localization by looking for locations with more than 100 cpm radiation after averaging the measurements over the last 270 days but *before interpolating the data*. This is to avoid that interpolation modifies the position of the hotspots, or even eliminates them. We show the results of detection when using the raw measurements (top left), and after the application of Release-GeoInd 30m (bottom left), GeoInd 300m (top right), and Rounding 2 (bottom right) in Figure 8. We see that noise-based mechanisms spread the measurements and, as the noise increases, create additional hotspots. Thus, these mechanisms are useless for hotspot detection: the results cannot be properly interpreted. Imagine a hotspot in a place known to present high radiation, thus being already closely monitored by the authorities. Finding such hotspot is not alarming. However, after spreading, the finding of hotspots conveys a much different message, especially when they appear in zones that had low radiation in the past.

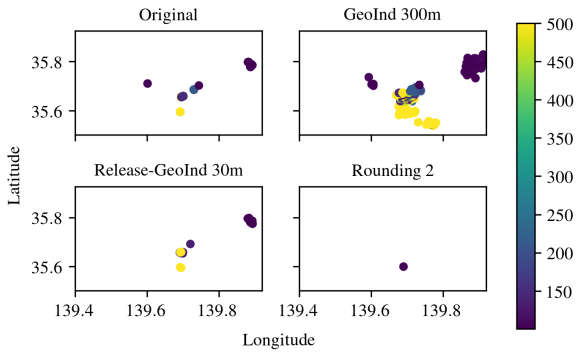

Figure 8: Safecast: Hotspot detection for areas with at least 100 cpm. Comparison of various defenses vs the original hotspots.

Generalization such as Rounding 2, which provides a good privacy-utility tradeoff for aggregated statistics, also performs poorly. In this case, the defense causes hotspots to disappear, potentially causing a dangerous situation if a high radiation zone is marked as safe. We also carry out experiments with hiding mechanisms and find that, similarly to Rounding, they miss some of the original hotspots.

Safecast takeaways. Considering only the privacy loss, GeoInd variants (except GeoInd 50m) and Rounding to 2 decimals seem to offer the best performance, while Random sampling and Release's protection is generally bad in terms of precision, and also too dependent on users' movement profiles. However, an analysis of the utility impact indicates that *none of the existent LPPMs is well suited* for the Safecast setting. The semantic interpretation results indicate that even if two defenses produce similar average results, the outliers they create can convey opposite messages. Furthermore, even a slight addition of noise or generalization can hinder the project's ability to correctly locate abnormal events. These limitations effectively prevent Safecast from deploying them to protect their users' privacy.[11]

[11] This statement was verified in communication with Safecast.

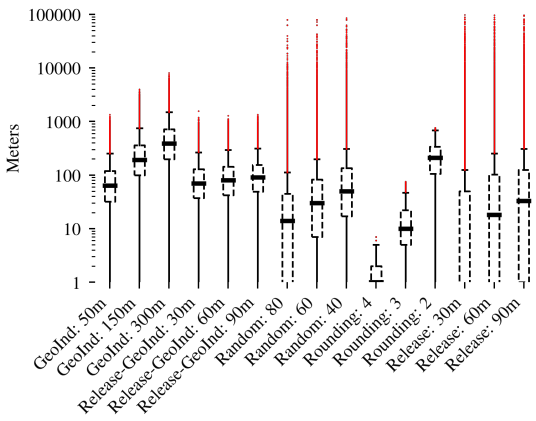

Figure 9: Radiocells: Utility loss (distance to tower location).

4.4.2 Radiocells

Radiocells' utility function is rather different than the one for Safecast. Instead of averaging measurements associated to a location, Radiocells averages all reported coordinates associated to an antenna to derive its position. We show the related utility loss for different LPPMs in Figure 9.

All GeoInd variants induce high utility loss, with medians between 80 and 400 meters, and with outliers beyond 2 kilometers. Surprisingly, in this use case hiding mechanisms (Release and Random) have many outliers. After manual inspection, we found out that several users had inconsistent measurements. For instance, a user was swapping her measurements' longitudes and latitudes in a random pattern. Other outliers are caused by providers moving their antennas IDs creating mixed measurements for a given ID. Furthermore, hiding defenses also influence the number of antennas located. In our dataset, we detect from 10.2% up to 18.6% fewer antennas when the Release defense is used, and the Random mechanism eliminates from 2.6% up to 13.7% of them.

The best mechanism in the Radiocells dataset is Release GeoInd which offers on average lower utility loss than other LPPMs and provides acceptable privacy. However, some antennas might be moved over a kilometer away. The next best alternative is Rounding 2 that has a higher median utility loss but no outliers. However, as the goal of the project is to *accurately* detect antennas in order to give individuals the ability to geolocate themselves offline or to enable scientific studies, a median error of 100 meters (Release GeoInd) or 200 meters (Rounding 2) is considered too large and precludes Radiocells from deploying them.

5 What's Next?

In this section, we elaborate on technical and non-technical steps to enhance privacy at smaller utility cost in the context of MCS applications.

5.1 Towards Effective Defenses

We first discuss possible strategies to improve the trade-off between users' privacy and MCS utility.

An unexplored approach is the use of advanced cryptographic protocols to compute the values of interest for MCS without revealing the users' individual values to the providers [41]. For instance, users could use multi-party computation to collaboratively compute aggregates and only report the result to the provider. However, cryptographic approaches require high computational power on the users' side and increase the bandwidth needs to perform the joint computation. Furthermore, this would limit the availability of raw measurements for analysis other than those predefined by the cryptographic protocols, which is at odds with the principles of open data and open science defended by most of the MCS platforms.

In our evaluation, we only considered spatial generalization. Another avenue to explore would be to also generalize the time dimension. On its own time obfuscation cannot hide patterns revealed by repeated visits. However, combined with full de-identification and hiding of users could reduce the inference power of the adversary. For instance, the MCS service provider could release a batch of measurements once a day or once a week without linking these to any user identifier. These techniques would be cheaper than the use of cryptography, but require trust on the service provider to properly apply sanitization and protect the raw data.

A third research path is the co-design of defenses and aggregation algorithms. In this paper, we have considered that the output of the LPPMs is directly input to the utility functions currently used by MCS providers. However, it would be possible that the providers adapt their data processing to account for noise, using statistical methods or machine learning, as done in fields that rely on noisy sensors [79] or train in different settings from which they are deployed [43, 72, 85].

Finally, MCS could provide users with dedicated local software (e.g., building on our evaluation method) to alert them regarding the privacy dangers of publishing raw location data. Such a system would allow them to selectively hide some of their measurements, reducing the confidence of inference attacks. We note that, when building such a tool, one would like to consider attacks beyond the POI-based inferences considered in this paper. For instance, it has been shown that co-locations can unveil social links [38, 44]. We run a preliminary evaluation to learn whether our MCS setting is also prone to such an attack. We identified 50 unique pairs of users with real names and at least one co-location (similar latitude, longitude, and time) in the Safecast dataset. We could validate 16 of these pairs as real friendships using information available on online social networks, i.e., yielding a 32% correct inference rate. Note that many of the other pairs could not be verified because either users were not part of any social network or they did not publicly reveal their social links. More advanced methods, such as measuring the amount of time

two users are co-located or the number of different locations where two users jointly report their locations [30,38,87] could further improve these results. Therefore, new defenses need to also obfuscate co-locations [71].

5.2 Privacy Considerations for Developers

In our study, we identified a number of issues related to the collection and sharing of data that, even though cannot fully prevent inference, could make inference attacks detectable and could render potential attackers accountable.

A first consideration to make is the type of policy under which MCS publish the collected data. While making large datasets available to everyone for unrestricted use is admirable, and certainly of high value for the academic community, it can have serious implications for the altruistic contributors. To reduce this risk, developers could add clauses to the policies that not only mandate that use of the data is properly acknowledged, but also that it is well documented, implying that researchers or other individuals have to disclose how they have processed the data, and for which purpose.

Second, both Safecast and Radiocells datasets are available for download without the need for authentication. This hinders traceability of who has the data, and thus enables stealthy attacks where nor the users neither the applications are aware of the danger. Like in other projects that make data available for research and other purposes (e.g., the Drebin project[12]), these sites could require simple registration to maintain a log of who has had access to the datasets. Together with the previous requirement, which would include documentation of sharing, it should help mitigate the risks.

Third, these applications typically do not perform any control on who are the contributors. This poses a particular problem when it comes to children. In many jurisdictions, children's data are subject to particular legislation [37,47], and in particular require the parents' consent to be collected and processed. The lack of control upon collection implies that the datasets could contain children's geo-located data collected illegally. Adding control would solve this problem and also support the previous two points.

Finally, the datasets we studied contain data from users from all over the world. These users, therefore, are subject to different legislations that regulate how their data can be processed. While this may not be a problem for corporations or criminals that want to exploit the datasets, it creates a hurdle for researchers who have to obtain approval from their institution for data processing. This problem arised during our discussions with our institution's Ethical Review Committee, and almost caused us to stop the project. In other words, lack of proper documentation may limit the free use of the data for science, effectively hindering one of the main goals of these applications. Better documentation as to the origin of data and its use possibilities would greatly facilitate the process.

[12] https://www.sec.cs.tu-bs.de/~danarp/drebin/

6 Related Work

We have covered the related work on LPPMs in Section 3.1 and the previous work on privacy quantification in Section 3.2. We complete this review of the literature with previous research on human mobility and its privacy implications.

Similar to [36, 49, 58, 59, 64, 66, 67, 69, 86], our POIs extraction attack is based on machine learning. Gambs and Killijian [52] also rely on POIs inference to build mobility Markov chains and de-anonymize traces. Gonzalez et al. [54] and Song et al. [82] study anonymized mobile phone data. Their results indicate that human trajectories have a high degree of temporal and spatial regularity, and that an individual's location data history is a unique identifier. De Montjoye et al. [39] investigate how the uniqueness of mobility traces decays depending on their resolution. They show that uniqueness cannot be avoided by lowering the resolution of a dataset. While these works aim at understanding the uniqueness of individuals or de-anonymize them, we focused on inferences that rely on labeled traces.

Similar to us, Gambs et al. [50] develop a platform for evaluating various sanitization methods and attacks on geo-located data. They focused on evaluating LBSs, while we evaluate the effectiveness of defenses on MCS applications. We also use different privacy metrics, and utility functions, tailored to the MCS scenario. Finally, Drakonakis et al. [42] explore the privacy loss stemming from by public location metadata. They propose a tool to infer users' regions of interest and, by experimenting with data gathered from Twitter, they illustrate the accuracy of their tool in pinpointing users' sensitive locations. Furthermore, they highlight how the spatial data provide additional context on the information shared by the user. We use similar techniques to prove that these inferences are also possible in MCS. For further information about the security and privacy landscape of location data we refer the reader to the surveys in [53, 65, 76, 84].

7 Conclusion

Mobile crowdsourcing is an increasingly popular way to collect geo-located data from millions of contributors. We present the first study on privacy implications of MCS applications. We study the applicability of well-established location privacy defenses created for LBSs. We show that neither the location privacy and utility metrics typically found in the literature nor the existing privacy-preserving mechanisms are well-suited for the MCS case. On the one hand, given the persistent patterns stemming from continuous collection, these solutions provide less privacy than in the case of LBSs where locations are revealed once. Second, the existing mechanisms are optimized to provide utility regarding the location of the users, but MCS applications rely on measurements associated to these locations, or on some function of the locations. Therefore, state-of-the-art defenses have a detrimental impact on

the MCS utility.

In conclusion, we identify an underexplored space in the location privacy literature, that is of practical relevance for many new applications. We have outlined promising lines to improve the situation. We hope that our findings spawn new research that soon enables the deployment of privacy-preserving crowdsourcing applications.

Acknowledgments

This work has been funded by the German Science Foundation (DFG) as part of the project A1 within the RTG 2050 "Privacy and Trust for Mobile Users".

References

[1] URL: https://www.spotteron.net/apps.

[2] URL: https://support.google.com/wifi/answer/6246642.

[3] URL: https://privacy.microsoft.com/en-us/windows-10-location-and-privacy.

[4] URL: https://location.services.mozilla.com.

[5] URL: https://www.openstreetmap.org.

[6] URL: https://opensignal.com/.

[7] URL: https://blog.safecast.org.

[8] URL: https://www.opencellid.org.

[9] URL: https://radiocells.org.

[10] URL: https://www.skyhookwireless.com.

[11] URL: https://www.sensorly.com.

[12] URL: http://www.cellumap.com.

[13] URL: https://www.mapillary.com.

[14] URL: https://play.google.com/store/apps/details?id=com.opensignal.weathersignal.

[15] URL: https://www.waze.com.

[16] URL: https://www.qualcomm.com/solutions/automotive/drive-data-platform.

[17] URL: https://www.gokamino.com.

[18] URL: http://www.app-store.es/stereopublic.

[19] URL: https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy/.

[20] URL: https://www.bellingcat.com/resources/articles/2018/07/08/strava-polar-revealing-homes-soldiers-spies/.

[21] URL: http://www.whosdrivingyou.org/blog/ubers-deleted-rides-of-glory-blog-post.

[22] URL: https://github.com/SpatialVision/differential_privacy.

[23] URL: https://en.wikipedia.org/wiki/Fukushima_Daiichi_nuclear_disaster.

[24] URL: https://github.com/spring-epfl/MCSAuditing.

[25] URL: https://developer.mapquest.com/documentation.

[26] URL: http://www.gqelectronicsllc.com/GMC_Safty_Guide.jpg.

[27] URL: http://earthpy.org/interpolation_between_grids_with_ckdtree.html.

[28] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *CCS*, 2013.

[29] Enrique Estellés Arolas and Fernando González-Ladrón-de-Guevara. Towards an integrated crowdsourcing definition. *J INF SCI*, 2012.

[30] Michael Backes, Mathias Humbert, Jun Pang, and Yang Zhang. walk2friends: Inferring social links from mobility profiles. In *CCS*, 2017.

[31] Bhuvan Bamba, Ling Liu, Péter Pesti, and Ting Wang. Supporting anonymous location queries in mobile environments with privacygrid. In *WWW*, 2008.

[32] Vincent Bindschaedler and Reza Shokri. Synthesizing plausible privacy-preserving location traces. In *IEEE S&P*, 2016.

[33] Konstantinos Chatzikokolakis, Ehab Elsalamouny, and Catuscia Palamidessi. Efficient utility improvement for location privacy. *PETS*, 2017.

[34] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. A predictive differentially-private mechanism for mobility traces. In *PETS*, 2014.

[35] Rui Chen, Gergely Ács, and Claude Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *CCS*, 2012.

[36] Sung-Bae Cho. Exploiting machine learning techniques for location recognition and prediction with smartphone logs. *NEUROCOMPUTING*, 2016.

[37] U.S. federal trade commission. complying with coppa: Frequently asked questions, 2015.

[38] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *P NATL ACAD SCI USA*, 2010.

[39] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *SCIENTIFIC REPORTS*, 2013.

[40] Deloitte. The three billion, enterprice crowdsourcing and the growing fragmentation of work. URL: https://www2.deloitte.com/content/dam/Deloitte/de/Documents/Innovation/us-cons-enterprise-crowdsourcing-and-growing-fragmentation-of-work%20(3).pdf.

[41] Daniel Demmler, Thomas Schneider, and Michael Zohner. Aby-a framework for efficient mixed-protocol secure two-party computation. In *NDSS*, 2015.

[42] Kostas Drakonakis, Panagiotis Ilia, Sotiris Ioannidis, and Jason Polakis. Please forget where i was last summer: The privacy risks of public location (meta) data. In *NDSS*, 2019.

[43] Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Rebecca S. Portnoff, Sadia Afroz, Damon McCoy, Kirill Levchenko, and Vern Paxson. Identifying products in online cybercrime marketplaces: A dataset for fine-grained domain adaptation. In *Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 2598–2607, 2017.

[44] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *P NATL ACAD SCI USA*, 2009.

[45] Hariton Efstathiades, Demetris Antoniades, George Pallis, and Marios D. Dikaiakos. Identification of key locations based on online social network activity. In *ASONAM*, pages 218–225. ACM, 2015.

[46] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[47] 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 Data Protection Regulation). *Official Journal of the European Union*, 2016.

[48] Kassem Fawaz and Kang G Shin. Location privacy protection for smartphone users. In *CCS*, 2014.

[49] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. Evaluating the privacy risk of location-based services. In *FC*, 2011.

[50] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show me how you move and i will tell you who you are. In *SPRINGL*, 2010.

[51] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *MPM*, 2012.

[52] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. De-anonymization attack on geolocated data. *TRUSTCOM*, 2013.

[53] Gabriel Ghinita. Privacy for location-based services. *Synthesis Lectures on Information Security, Privacy, & Trust*, 2013.

[54] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 2008.

[55] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *MobiSys*, 2003.

[56] Nicolas Haderer, Romain Rouvoy, Christophe Ribeiro, and Lionel Seinturier. Apisense: Crowd-sensing made easy. *ERCIM News*, 2013.

[57] Wajih Ul Hassan, Saad Hussain, and Adam Bates. Analysis of privacy protections in fitness tracking social networks-or-you can run, but can you hide? In *USENIX*, 2018.

[58] Min-Oh Heo, Myung-Gu Kang, Byoung-Kwon Lim, Kyu-Baek Hwang, Young-Tack Park, and Byoung-Tak Zhang. Real-time route inference and learning for smartphone users using probabilistic graphical models. *Journal of KIISE*, 2012.

[59] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE PERVAS COMPUT*, 2006.

[60] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Preserving privacy in gps traces via uncertainty-aware path cloaking. In *CCS*, 2007.

[61] Leping Huang, Hiroshi Yamane, Kanta Matsuura, and Kaoru Sezaki. Silent cascade: Enhancing location privacy without communication qos degradation. In *SPC*.

[62] Huan Feng Kassem Fawaz and Kang G Shin. Anatomization and protection of mobile apps' location privacy threats. In *USENIX*, 2015.

[63] Youssef Khazbak and Guohong Cao. Deanonymizing mobility traces with co-location information. In *CNS*, 2017.

[64] John Krumm. Inference attacks on location tracks. In *PERVASIVE*, 2007.

[65] John Krumm. A survey of computational location privacy. *PERS UBIQUIT COMPUT*, 2009.

[66] L. Liao, D. Fox and H. Kautz. Learning and inferring transportation routines. *AAAI*, 2004.

[67] Lin Liao, Dieter Fox, and Henry Kautz. Location-based activity recognition. In *NIPS*, 2006.

[68] Changsha Ma and Chang Wen Chen. Nearby friend discovery with geo-indistinguishability to stalkers. *FNC/MobiSPC*, 2014.

[69] Wesley Mathew, Ruben Raposo, and Bruno Martins. Predicting future locations with hidden markov models. In *UbiComp*, 2012.

[70] Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, Mathias Humbert, and Jean-Pierre Hubaux. Quantifying interdependent privacy risks with location data. *TMC*, 2017.

[71] Alexandra-Mihaela Olteanu, Mathias Humbert, Kévin Huguenin, and Jean-Pierre Hubaux. The (co-)location sharing game. In *PoPETs*, 2019.

[72] Rebekah Overdorf and Rachel Greenstadt. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *PoPETs*, 2016(3):155–171, 2016.

[73] Simon Oya, Carmela Troncoso, and Fernando Pérez-González. Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms. In *CCS*, 2017.

[74] Dau Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, 2000.

[75] Layla Pournajaf, Li Xiong, Vaidy Sunderam, and Xiaofeng Xu. Stac: Spatial task assignment for crowd sensing with cloaked participant locations. In *SIGSPATIAL/GIS*, 2015.

[76] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. The long road to computational location privacy: A survey. *IEEE Commun. Surv. Tutor.*, 2018.

[77] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. *NDSS*, 2018.

[78] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. What does the crowd say about you? evaluating aggregation-based location privacy. *PETS*, 2017.

[79] Jing Shi, Rui Zhang, Yunzhong Liu, and Yanchao Zhang. Prisense: privacy-preserving data aggregation in people-centric urban sensing systems. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[80] Reza Shokri, Julien Freudiger, Murtuza Jadliwala, and Jean-Pierre Hubaux. A distortion-based metric for location privacy. In *WPES*, 2009.

[81] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying location privacy. In *IEEE S&P*, 2011.

[82] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *SCIENCE*, 2010.

[83] Fung Global Retail & Technology. Crowdsourcing:seeking the wisdom of crowds. URL: http://www.deborahweinswig.com/wp-content/uploads/2016/07/Crowdsourcing-Report-by-Fung-Global-Retail-Tech-July-12-2016.pdf.

[84] Manolis Terrovitis. Privacy preservation in the dissemination of location data. *SIGKDD Explorations*, 2011.

[85] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote sensing magazine*, 4(2):41–57, 2016.

[86] Jorim Urner, Dominik Bucher, Jing Yang, and David Jonietz. Assessing the influence of spatio-temporal context for next place prediction using different machine learning approaches. *ISPRS INT GEO-INF*, 2018.

[87] Hongjian Wang, Zhenhui Li, and Wang-Chien Lee. Pgt: Measuring mobility relationship using personal, global and temporal factors. In *ICDM*, 2014.

[88] Yonghui Xiao and Li Xiong. Protecting locations with differential privacy under temporal correlations. In *CCS*, 2015.

[89] Hui Zang and Jean Bolot. Anonymization of location data does not work: A large-scale measurement study. In *MobiCom*, 2011.

A Appendix

A.1 Density Based Clustering (DBSCAN)

The algorithm receives as input all locations (also referred to as points) reported by a user, the minimum required amount of points per cluster, and the maximum allowed distance between the cluster's points. It outputs a label for every point, indicating to which cluster it belongs, or if it has been labeled as noise.

DBSCAN starts by randomly selecting a point c. Then, it finds all points p that are in distance ε from this point. Then, from the points p reachable from the first point, it tries to find more points q where q are reachable directly from p but not from c. If at the end of this procedure the minimum points have not been reached, it moves to another random point and starts all over again. In order to use our locations which are in latitudes and longitude, we converted the distance ε to radians first. Moreover, we used a ball tree data structure to speed up the neighbors queries.

A.2 Geo-Indistinguishability

The noise is drawn by first transforming the location to polar coordinates. Then, the angle is drawn randomly between 0 and 2π while the distance is drawn from

$$C^{-1}(\rho) = -\frac{1}{\varepsilon}(W^{-1}(\frac{\rho-1}{e})+1)$$

with W^{-1} denoting the -1 branch of the Lambert W function. Finally, the generated distance and angle are added to the original location.

A.3 Optimal Remapping

For the optimal remapping technique we follow these steps; For performance reasons, we first round each location to 3 digits, in order to merge nearby locations together. Then, we calculate the probability of each coordinate. Afterwards, we convert all coordinates to a Cartesian system using their distance from the center of the Earth. A useful tutorial on this can be found in [27]. Using the Cartesian coordinates we build a KD-Tree for efficient nearest neighbor calculations. Then, for every location where GeoInd has been applied, we query all nearest neighbors in a region r'. This r' is set to be as the 99% percentile of the distribution that generated the parameter r used in GeoInd. In other words, the user has 99% chance of being remapped somewhere within this distance. For all neighboring points, we compute the posterior and then, we calculate the geometric median of those coordinates using the iterative Weiszfeld's algorithm. The geometric median minimizes the average Euclidean distance and hence, returning us the new, optimal (in terms of utility as privacy should remain the same) location.

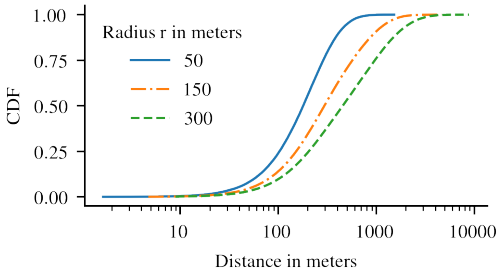

Figure 10: GeoInd noise magnitude for different radius ($l = \ln(1.6)$).

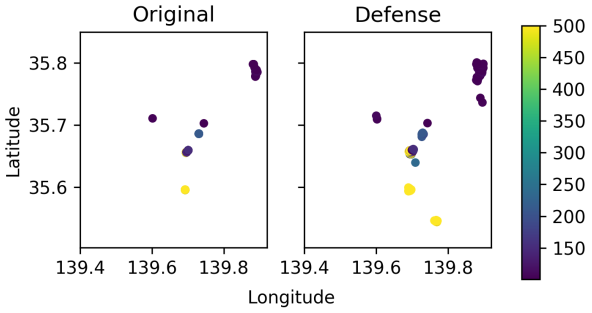

Figure 11: Safecast: Hotspot detection for areas with at least 100 cpm. The presented defense is GeoInd with 50m parameter.

A.4 Defenses evaluation

We include three more figures to complement the defense evaluation:

- Figure 10 portrays the CDF of the noise added by GeoInd. This noise is added on all GeoInd variants (Optimal Remapping and Release-GeoInd) and it is controlled by either the radius (r) or the privacy parameter (l).
- Figure 11 illustrates the hotspot detection results when GeoInd 50m is used. Even a slight addition of noise spreads the locations, not allowing Safecast to accurately detect elevated radiation regions.
- In Figure 12 we present the privacy gain results for each of the defense mechanisms for the whole Safecast dataset.

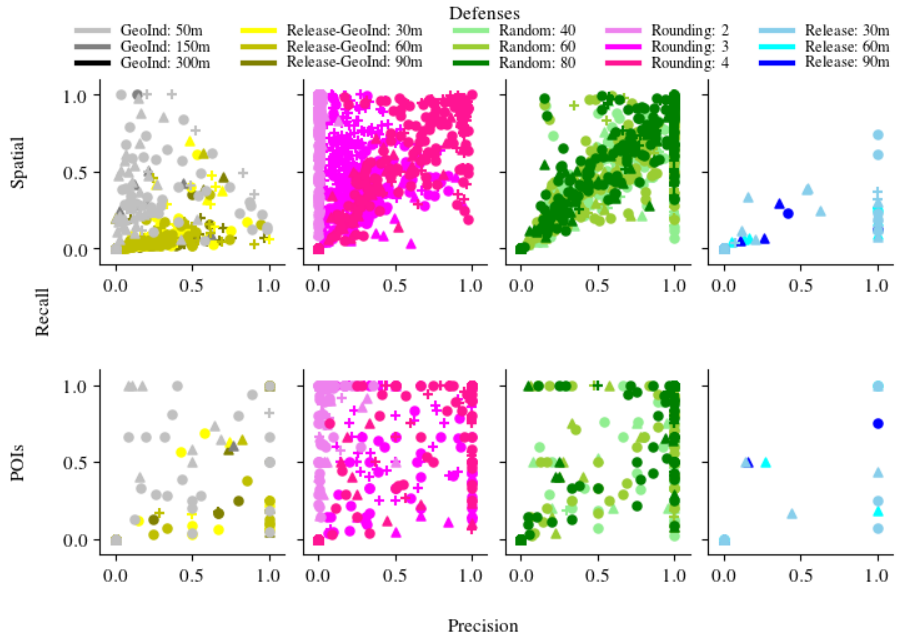

Figure 12: Safecast: Privacy gain for each of the defense mechanism (whole dataset).

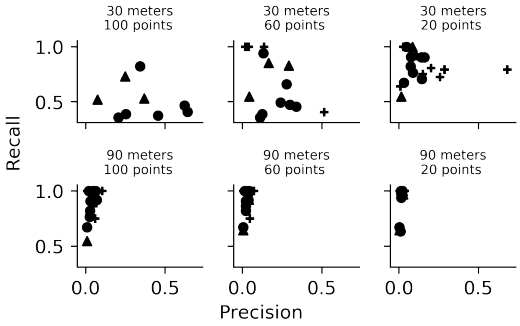

Figure 13: Precision and recall vs. clustering parameters for GeoInd ($r = 50$m) in Tokyo.

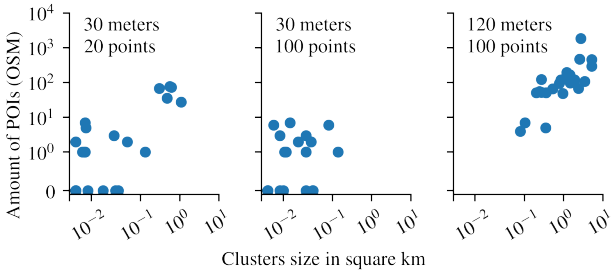

Figure 14: Clusters' size and amount of POIs per cluster vs. clustering parameters with GeoInd ($r = 50$m) in Tokyo.

A.5 Experimental results

Adjusting the clustering parameters. We now study the influence of the DBSCAN clustering parameters on our results. We show the difference in precision and recall for GeoInd (r=50 meters) when we vary both the maximum distance and the minimum number of point per cluster in Figure 13. As we increase the maximum distance between points and decrease the minimum required points per cluster, the results concentrate on the upper left corner of the diagram. This is because as the parameters become 'looser', the resulting clusters grow in size increasing recall (more likelihood of covering all users' original clusters) but reducing precision due to many false positives. Furthermore, increasing the cluster size increases the adversary's cost, as the clusters contain a larger number of POIs (Figure 14) which requires more filtering and increases the probability of having false positives.