Show Me How You Move and I Will Tell You Who You Are

Sébastien Gambs
Université de Rennes 1 - INRIA / IRISA
IRISA, Campus de Beaulieu
Avenue du Général Leclerc
35042 Rennes Cedex, France
sgambs@irisa.fr

Marc-Olivier Killijian
CNRS ; LAAS ; 7 avenue du Colonel Roche, F-31077 Toulouse, France
Université de Toulouse ; UPS , INSA , INP, ISAE ; LAAS
marco.killijian@laas.fr

Miguel Núñez del Prado Cortez
CNRS ; LAAS ; 7 avenue du Colonel Roche, F-31077 Toulouse, France
Université de Toulouse ; UPS , INSA , INP, ISAE ; LAAS
mnpc@computer.org

ABSTRACT

Due to the emergence of geolocated applications, more and more mobility traces are generated on a daily basis and collected in the form of geolocated datasets. If an unauthorized entity can access this data, it can used it to infer personal information about the individuals whose movements are contained within these datasets, such as learning their home and place of work or even their social network, thus causing a privacy breach. In order to protect the privacy of individuals, a sanitization process, which adds uncertainty to the data and removes some sensible information, has to be performed. The global objective of GEPETO (for *GEoPrivacy Enhancing TOolkit*) is to provide researchers concerned with geo-privacy with means to evaluate various sanitization techniques and inference attacks on geolocated data. In this paper, we report on our preliminary experiments with GEPETO for comparing different clustering algorithms and heuristics that can be used as inference attacks, and evaluate their efficiency for the identification of point of interests, as well as their resilience to sanitization mechanisms such as sampling and perturbation.

Categories and Subject Descriptors

D.4.6 [**Operating Systems**]: Security and Protection; K.4.1 [**Computers and Society**]: Public policy issues—*privacy*; H.2.8 [**Database Applications**]: Spatial databases and GIS

General Terms

Security

Keywords

Privacy, Geolocated data, Geo-privacy, Inference attacks, Sanitization, Clustering.

1. INTRODUCTION

A *geolocated system* is an object or device which has an associated location. For instance, it can be a smartphone or a GPS-equipped vehicle. Usually, a geolocated system belongs to an individual (or to a group of individuals, such as a family) and as such its location corresponds to the location of its owner(s). Geolocated data is already publicly available and sometimes easy to obtain. For instance, some persons diffuse publicly, almost in real-time, their current location via social application such as Twitter which in turn can be collected to predict whether or not they are currently at home[1]. Other applications, such as Google Latitude[2], allow to track the movements of friends' cellphones and display their position on a map. Apart from these social applications, there are also other public sources of information that can be exploit by a potential adversary for causing a privacy breach, such as free and easy access to geographic knowledge with Google Maps[3], Yahoo!Maps[4] and Google Earth[5].

We have started to explore, study and axiomatize the different types of inference attacks on geolocated data and basically our main finding is that among all the *Personal Identifiable Information* (PII), learning the location of an individual is one of the greatest threat against his privacy. For instance, the spatio-temporal data of an individual can be used to infer the location of his home and workplace, to trace his movements and habits, to learn information about his center of interests or even to detect a change from his usual behaviour. We provide a brief overview and classification of inference attacks on geolocated data in Section 2.

One of the main challenge for geoprivacy is to balance the benefit for an individual of participating to a geolocated application with the privacy risks he incurs by doing so. For example, if Alice's car is equipped with a GPS and she accepts to participate to the real-time computation of the traffic map, this corresponds to a task that is mutually beneficial to all the drivers but at the same time Alice wants to have some privacy guarantees that her individual locations will be protected and not broadly disclosed. In practice, we clearly advocate to follow the "privacy by design" paradigm which explicitly takes into account the privacy issues in the design process of a geolocated application, rather than simply de-

[1] http://pleaserobme.com/
[2] http://www.google.com/latitude
[3] http://maps.google.com/
[4] http://maps.yahoo.com/
[5] http://earth.google.com/

SPRINGL '10 November 2, 2010, San Jose, CA, USA

ploying it and wait for the possible disastrous consequences.

We emphasize that simply removing the identifiers of individuals or replacing them by a pseudonym is usually not sufficient to protect their privacy. Instead, a *sanitization* process, which adds uncertainty to the data and removes some sensible information, has to be performed. This loss of data, incurred by the sanitization process, comes with a dilemma: it certainly brings some privacy guarantees but at the cost of a decrease of utility due to the quality degradation of the data. Therefore, there is often a trade-off between the utility of the global task and the privacy protection of individuals. In Section 3, we describe some sanitization algorithms and methods for preserving geoprivacy before reporting in Section 4 our ongoing work on GEPETO (for *GEoPrivacy Enhancing TOolkit*) [1], a flexible open source software which can be used to visualize, sanitize, attack and measure the utility of a particular geolocated dataset. Finally, we conclude with a brief discussion in Section 5.

2. INFERENCE ATTACKS ON GEOLOCATED DATA

An *inference attack* is an algorithm that takes as input some geolocated data D, possibly with some auxiliary information aux, and produces as output some additional knowledge. For example, an inference attack may consist in identifying the house or the place of work of an individual. The auxiliary information reflects any *a priori* knowledge that the adversary might have gathered (for instance through previous attacks and by accessing some public data source) and which may help him in conducting an inference attack. We propose to classify the inference attacks according to (at least) three dimensions such as the type of data it works on, the objective of the attack as well as the specific technique used.

2.1 Geolocated Data

Nowadays, the rapid growth and development of geolocated applications has multiplied the potential sources of geolocated data. The geolocated data generated by these diverse applications varies in its exact form and content but it also shares some common characteristics. Regarding the type of data, we differentiate mainly between mobility traces and contact traces. A *mobility trace* is characterized by:

- An *identifier*, which can be the real identifier of the device (e.g. "Alice's phone"), a pseudonym or even the value "unknown" (when full anonymity is desired). A pseudonym is generally used when we want to protect the true identity of the system while still being able to link different actions performed by the same user.
- A *spatial coordinate*, which can be a GPS position (e.g. latitude and longitude coordinates), a spatial area (e.g. the name of a neighbourhood in a particular city) or even a semantic label (e.g. "home" or "work").
- A *time stamp*, which can be the exact date and time or just an interval (e.g. between 9AM and 12AM).
- Additional information such as the speed and direction for a vehicle, the presence of other geolocated systems or individuals in the direct vicinity or even the accuracy of the estimated reported position. For instance, some geolocated systems are able to estimate the precision of their estimated location as a function of the number of GPS satellites they are able to detect.

Contact traces are a specific form of mobility traces which consist in the recording of encounters between different devices. This kind of trace is composed of the identifiers of the devices and a time stamp. It may be recorded for instance by a device which has no integrated capacity for geopositioning but is capable of probing his neighbourhood to detect the presence of other devices (e.g. using Bluetooth neighbour discovery).

A *geolocated dataset* D is a dataset which contains mobility traces of individuals. Technically, this data may have been collected either by recording locally the movements of each geolocated system for a certain period of time, or centrally by a server which can track the location of these systems in real-time. A *trail of traces* is a collection of mobility traces that corresponds to the movements of an individual over some period of time. A geolocated dataset D is generally constituted by an ensemble of trails of traces from different individuals. The Crawdad project[6] is an example of a public repository giving access to geolocated datasets, which can be used for research purpose.

2.2 Objective of the Attack

An adversary attacking some geolocated data may have various objectives ranging from identifying the home of the target to reconstructing his social network, through obtaining knowledge of his favourite jogging tracks. More precisely, the objective of an inference attack may be to:

- *Identify important places*, called *Points Of Interests* (POIs), which characterize the interests of an individual [2]. A POI may be for instance the home or place of work of an individual or locations such as a sport center, theater or the headquarters of a political party. Revealing the POIs of a particular individual is likely to cause a privacy breach as this data may be used to infer sensible information such as hobbies, religious beliefs, political preferences or even potential diseases. For instance, if an individual has been visiting a medical center specialized in a specific type of illness, then it can be deduced that he has a non-negligible probability of having this disease.
- *Predict the movement patterns of an individual* such as his past, present and future locations. From the movement patterns, it is possible to deduce other PII such as the mode of transport, the age or even the lifestyle. According to some recent work [3, 4], our movements are easily predictable by nature. For instance, the authors of these papers have estimated to 93% the chance of correctly guessing the future location of a given individual after some training period on his mobility patterns.
- *Link the records of the same individual*, wich can be contained in different geolocated datasets or in the same dataset, either anonymized or under different pseudonyms. This is the geoprivate equivalent of the *statistical disclosure risk* where privacy is measured according to the risk of linking the record of the same

[6] `http://crawdad.cs.dartmouth.edu/`

individual in two different databases (e.g., establishing that a particular individual in the voting register is also a specific patient of an hospital [5]). In a geolocated context, the purpose of a linking attack might be to associate the movements of Alice's car (contained for instance in dataset A) with the tracking of her cell phone locations (recorded in another dataset B). As the POIs of an individual and his movement patterns constitute a form of fingerprinting, simply anonymizing or pseudonymizing the geolocated data is clearly not a sufficient form of privacy protection against linking attacks. For example, Colle and Kartridge [6] have shown that even the pair home-work becomes almost unique per individual, and thus acts as a quasi-identifier, if the granularity is not coarse enough (e.g., if the street is revealed instead of the neighbourhood).

- *Discover social relations* between individuals by considering for instance that two individuals that are in contact during a non-negligible amount of time share some kind of social link (of course false positive may happen) [7]. This information can also be derived from mobility traces by observing that certain individuals are in the vicinity of each other on a frequent basis.

2.3 Inference Technique

We describe thereafter some learning algorithms and methods that can be used as inference technique:

- *Clustering* is a form of unsupervised learning that tries to group objects that are similar in the same cluster while putting objects that are dissimilar in different clusters. A clustering algorithm needs a *distance measure* (or a similarity metric) to quantify how far/similar are two objects relative to each other and to drive the clustering process. A natural distance between two locations is simply the Euclidean distance but of course more complex metrics can be used, such as the length of the shortest path according to the existing roadmap. For instance, k-means is an iterative clustering algorithm that outputs k clusters as well as their respective centres (which are effectively the average of the locations within each cluster). This algorithm can be used straightforwardly to discover the POIs of one particular individual if it is fed only with his data [8], or the generic *hotspots* if it is given the geolocated data of a whole population. Hoh, Gruteser, Xiong and Alrabady have performed a study [9] on the geolocated data of vehicles within the Detroit area (Michigan, USA). The goal of their study was to automatically discover the home of the vehicles' drivers. The authors have used a clustering algorithm to automatically identify the houses and their findings is that among the 2 neighbourhoods and the 65 persons on which the authors have focused, the estimated houses correspond at 85% to the houses that a human would have recognized[7]. More complex techniques such as density-based clustering [10] can be used instead of k-means to overcome some of its shortcomings, such as k the predefined number of clusters and the constraint that the shape of the clusters has to be spherical.

- *Probabilistic models* can be learned from the geolocated data of individuals, and then used either to identify them among a geolocated dataset (even when they are anonymous) or to predict their next movements. For instance, Lio, Fox and Kautz [11] have shown that it is possible to train a relational Markov network, so that it can predict with a relatively good accuracy the next location of an individual or his current activities. Another possibility is to use an algorithm for tracking the movement of targets [12] to reconstruct the paths followed by several individuals in a geolocated dataset even if they are anonymous.

- *Heuristics* gives also good results in practice [13] for identifying POIs at a relatively low cost. An heuristic can be as simple as choosing the last stop before midnight or the average (or median) of several stop locations for identifying the home or the most stable location during the day for finding the place of work.

- *Data coming from social applications* is a possible source of information that the adversary might draw on to attack the privacy of individuals. The website "Please Rob Me" is a stricking example of how it is possible from publicly available information in the form of Twitter's posts (i.e. *tweets*) to build a classifier that can predict whether or not somebody is currently at home. Another example of social application is Google Latitude that offers the possibility of following in real-time on a map the movements of siblings and friends who have previously agreed to this service by confirming this on a SMS received on their phone[8]. However, some social applications such as Locaccino[9] tries to integrate explicitly the privacy issues in their design, by giving the possibility to a user to choose how he wants to disclose and share its location with its friends, and helping him understand what are the potential privacy risks that he might incur.

- *Data coming from public sources* is also a potential source of knowledge that can be exploited by the adversary. For instance, by using Google Maps and Yahoo!Maps the adversary can easily reconstruct the path followed by an individual between two consecutive mobility traces. Moreover, *reverse geocoding* tools exist that can transform a spatial coordinate into a physical address, which in turn can be cross-referenced with the corresponding entries in the Yellow Pages.

3. GEO-SANITIZATION MECHANISMS

A *sanitization algorithm* S takes as input a geolocated dataset D, introduces some uncertainty and removes some

[7] As the exact identity of the drivers have been kept secret it was not possible for the authors to compare directly the houses returned by the algorithm against the ground truth (i.e. the exact address of the drivers) which explained why this particular evaluation method was chosen.

[8] An infamous use of Google Latitude is known as the *shower attack* where a suspicious husband waits for his wife to take her shower, before sending the Google Latitude SMS to her cellphone, accepting this service on her behalf on the cellphone and then erasing it thus leaving not clue for her that she is now tracked.

[9] `http://locaccino.org/`

information from this dataset to increase the privacy of individuals whose movements are contained in the dataset. S produces as output D', a sanitized version of the original dataset D. The main idea behind sanitization is that, for a potential adversary, breaching the privacy of a particular user is harder when working on D' than with D. A sanitization procedure usually comes with some privacy guarantees. For instance, it can guarantee that at each time step, there is a minimum number of individuals in each spatial area. Possible sanitization techniques include:

- *Pseudonymization* replaces the common identifier of several mobility traces by either a randomly generated pseudonym (thus providing anonymity but not unlinkability) or by the *unknown* value (thus theoretically granting full anonymity and unlinkability)[10]. Pseudonymization is generally performed as the first step of a sanitization process but as such is often not sufficient for protecting the privacy of individuals.

- *Perturbation* methods [15] modify the spatial coordinate of a mobility trace by adding some random perturbation. For example, this noise can be generated uniformly or using Gaussian noise within a sphere of radius r centered on the original coordinate. If the geography of the surrounding area is not taken into account, it may happen that the perturbed coordinate corresponds to a location which has no physical sense (e.g., in the middle of a river or on a cliff).

- *Aggregration* merges several mobility traces into a single spatial coordinate. For instance, this spatial coordinate can be a surrounding spatial area such as a neighbourhood or the average of the mobility traces. During data preprocessing, a *clustering algorithm* (such as k-means) can be used to group traces that are close together into the same cluster while putting traces that are significantly distant into distinct clusters. This can be used to detect which traces should be merged together during an *aggregation* step. Another possibility is to detect traces occupying the same spatial area (for instance the same neighbourhood) at a certain moment in time and to replace each one of these individual traces by the coordinate of this spatial area.

- *Sampling* can be seen as a form of temporal aggregation. A *sampling* mechanism summarizes several mobility traces into fewer traces, generally by representing an ensemble of traces, which have occured within some time window, into one median or average trace. By decreasing the total number of traces, sampling has the additional benefit that it compresses the data and, therefore, reduces the computational resources needed to further sanitize the data.

- *Spatial cloaking* [16] is an extension of the concept of k-anonymity [5] to the spatio-temporal domain and a form of aggregation. The main idea is to ensure that at each time step, each individual is located within a spatial area that is shared by a least $k-1$ other individuals. This spatial area is disclosed instead of the exact location of these individuals, thus guaranteeing that even if an adversary can target the group where an individual is located, his behaviour will be indistinguishable from at least $k-1$ other individuals (k is a privacy parameter of the algorithm). A possible approach to achieve the property of spatial cloaking is to split recursively the space into areas of different sizes, until each area contains at least k individuals.

- *Mix-zones* [17] are inspired from the concept of mix-nets due to Chaum. Mix-zones are spatial areas where (1) no measurements about the locations of individuals are performed and (2) such that each individual entering a mix-zone will have a different pseudonym when he exits the mix-zone. The main purpose of a mix-zone is to make it more difficult to link the different actions of an individual. Areas or buildings with a high traffic are usually good candidates for mix-zones.

- *Swapping* consists in exchanging the mobility traces of two different individuals for a certain period of time. For example, by swapping Alice's and Bob's traces during one day, their behaviours become more atypical and less predictable.

- *Removing* the mobility traces that are deemed too sensible can also be considered as a sanitization procedure. In the same spirit, it is also possible to *add fake records* (called *dummies*) [18] inside the sanitized dataset to blend the true movements of individuals inside artificial data.

As sanitization leads to a loss of information, it is important to have a *utility metric* in order to compare the utility of the original dataset D and the sanitized one D'. The utility measure can either be generic, for instance linked to some global statistical properties of the dataset, or application-dependent, in which case it evaluates how well a particular application can be performed by using D' instead of D.

4. EXPERIMENTAL RESULTS

In this section, we report on our preliminary experiments towards building a generic toolkit for evaluating both sanitization methods and inference attacks on geolocated data. In particular, we describe some clustering algorithms and heuristics, which can be used as inference attack, and evaluate their efficiency for the identification of POIs, even after the application of sanitization mechanisms such as sampling and perturbation.

4.1 GEPETO

The global objective of GEPETO (for *GEoPrivacy Enhancing TOolkit*) [1] is to provide researchers concerned with geo-privacy with means to evaluate various sanitization techniques and inference attacks on geolocated data. GEPETO has an interface for the management of geolocated data and offers several ways to manipulate this data such as sanitization mechanisms, inference attacks and a visualisation tool to display this data on a world map. The main idea is to offer a generic and flexible tool so that anyone can easily plug a new sanitization technique or inference attack. Moreover,

[10] *Anonymity* can be defined as being able to perform a particular action without having to reveal his identity whereas *unlinkability* is a stronger notion that involves not being able to link two different actions that have been performed by the same user. Typically, performing different actions under a pseudonym (instead of using his real name) provides anonymity but not unlinkability. See [14] for more details.

the utility and visualization components provide means to evaluate the benefits of sanitization with regard to the success of inference attacks. To the best of our knowledge, there is almost no previous work that have tried to integrate all these features into a unified approach, with the exception of tools developed within the GeoPKDD project (and now the subsequent MODAP project) [19]. Another notable exception is [20] that tries to model formally the knowledge of the adversary with respect to the locations of individuals, and the possible counter-measures that these individuals might apply.

For the sake of demonstration, we begin by illustrating how GEPETO can be easily used to infer some private data about the taxi cabs of San Francisco, such as their home address for example. This geolocated data is available on the Crawdad repository. At first, GEPETO can simply be used to visualize the various mobility trails, and to characterise the geolocated data. When visualizing the data on the San Francisco map, one can easily recognize some hotspots, such as the San Francisco International Airport or various train and taxi stations. These hotspots being places where the taxis usually wait for customers during some period of time, many traces correspond to plots on these spots. GEPETO can thus be used to "manually" perform inference attacks by visualizing and mining geolocated data.

4.2 Playing with Heuristics

The first step was to explore the use of heuristics. For instance, by considering that the beginning and ending locations of the taxis, for each working day, might convey some meaningful information. This is the purpose of the *begin and end location finder* inference attack [1]. This attack is a simple heuristic assuming that the first and last recorded locations in a working day correspond to the departure and arrival points from a POI. The intuition is that when there is no mobility trace measured during a period longer than a given time threshold τ, this means that the individual had a "mobility break" and the place where he took this break is likely to be a POI. If τ is chosen sufficently large (e.g., 6 hours), this POI may be the home of an individual where he went to sleep after his work. First, we parse the mobility trails by looking for such breaks and extract the mobility traces that occured right before and right after.

We must say that this attack has been very fruitful. A first interesting inference was the identification of location of the taxi company main parking. Indeed, many cabs depart and arrive to this location after their working day, as they park their cab at the company headquarters. We were able to formally verify this statement simply by using the San Francisco Yellow Pages. The second category of statements that could be inferred from this attack directly concerns private information of the individual taxi drivers[11]. During this study, we examined the trails of 90 individual taxis chosen at random in the dataset. We used GEPETO to visualize the data of these 90 taxis after applying the *begin and end location finder* inference attack, manually picking those whose geolocated data seemed the most vulnerable. For 20 of these 90 taxis, the visualization of the resulting data result in a narrow neighbourhood for their homes with a pretty high confidence. Note however that, as we do not have the real addresses of the taxis, we were unable to formally validate these statements. However, we were able to some public data sources, such as Google Maps and Street View, to validate some of the inferred data. Indeed, for 10 of the 90 taxis checked, the attack resulted in an address (or a small portion of a street) where the taxi was parked during most of the breaks. This address is most probably the home address of the taxi driver. Figure 1 shows the result of a successful attack, together with a Google Maps view and a StreetView of the address. For the remaining 70 taxis examined, the *begin and end location finder* inference attack simply already identified hotspots (taxi stations, ...) that were already known.

[11]It is worth noting that for protecting their privacy, we blurred their address. However, the interested reader can obviously find the actual information by applying the same algorithms we did on the original dataset.

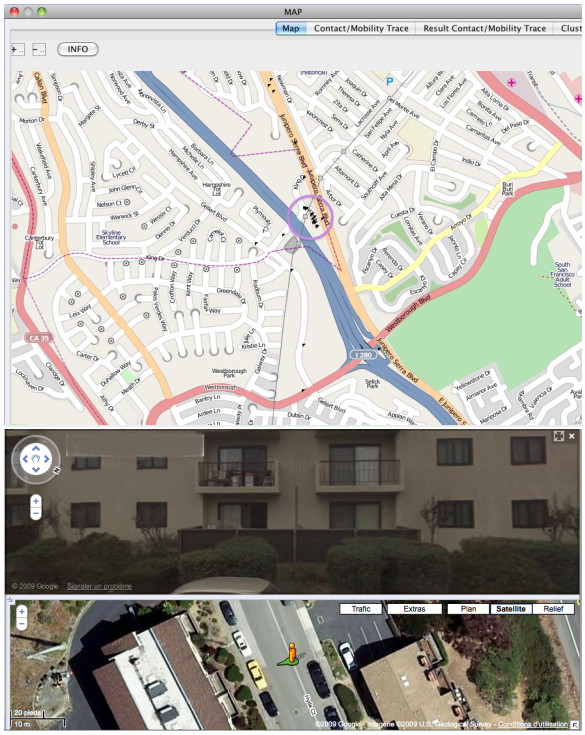

Figure 1: A successful *begin and end location finder* inference attack.

4.3 Experimenting with Clustering

The next step was to implement two clustering algorithms described in the litterature (Density-Joinable Cluster [21] and Density-Time Cluster [22]) and then to compare them with the Begin-end heuristic and to our own novel clustering algorithm (GEPETO clustering).

4.3.1 Description of the Clustering Algorithms

We describe thereafter succinctly the clustering algorithms that we have evaluated during our experiments.

- *Density-joinable cluster* (DJ Cluster) [21] is a clustering algorithm taking as input a minimal number of points *minpts*, a radius r and a trail of mobility traces M. This algorithm works in three phases. First, the preprocessing phase discards all the moving points (where speed is above zero) and then, squashs series of repeated static points into a single occurence for

each series. Then, the second phase clusters the remaining points based on neighbourhood density. More precisely, the number of points in the neighbourhood must be equal or greater than *minpts* and these points must be within radius r from the centroid of a set of points. Finally during the last phase, the algorithm merges the clusters which share at least one common point.

- *Density-time cluster* (DT Cluster) [22] is an iterative clustering algorithm taking as input a distance threshold d, a time threshold t and a trail of mobility traces M. First, the algorithm starts by building a cluster C composed of all the consecutive points within distance d from each other. Afterwards, the algorithm checks if the accumulated time of mobility traces within range is greater than the threshold t and created a cluster added to the list of POIs outputted if it is the case. Finally as a post-processing step, DT Cluster merges the clusters whose centroids are less than $d/3$ far from each other.

- *GEPETO cluster* is a novel clustering algorithm inspired from DT Cluster, which takes as input parameters a radius r, a time window t, a tolerance rate τ, a distance threshold d and a trail of mobility traces M. The algorithm starts by building iteratively clusters from a trail of traces M from mobility traces that are located within the time window t. Afterwards, for each cluster, if a fraction of the points (above the tolerance rate τ) are within radius r from the centroid, the cluster is integrated to the list of clusters outputted, whereas otherwise it it is simply discarded. Finally, as for DT Cluster, the algorithm merges the clusters whose centroids are less than d far from each other. See Algorithm 1 for a brief description of this method.

Algorithm 1 GEPETO clustering algorithm

Require: Trail of (mobility) traces M, time window t, radius r, tolerance rate τ, distance threshold d
Initialize N has being the number of records in the trail of traces M (i.e. $M.length$) and $cumulTime = 0$
Set L, the list of POIs found, has being the empty list
Create a empty cluster C
for $i = 0$ tol $N - 1$ **do**
  $cumulTime = cumulTime + (M[i+1].time - M[i].time)$
  if $cumulTime \leq t$ **then**
    Add the mobility trace $M[i]$ to cluster C
  else
    Compute the centroid of C
    $nbPtsOutsideRadius = 0$
    for $j = 0$ to $C.nbPts$ **do**
      if $distance(C[j], C.centroid) > r$ **then**
        $nbPtsOutsideRadius = nbPtsOutsideRadius + 1$
      end if
    end for
    if $nbPtsOutsideRadius/totalPoints < \tau$ **then**
      Add the cluster C to L
    end if
    Reset $cumulTime$ to 0 and create a new empty cluster C
  end if
end for
Merge clusters of L whose distance between centroids is less than d
return L, the list of POIs discovered (which are effectively the centres of the clusters)

4.3.2 Comparison and Resilience to Sanitization

These four algorithms (the Begin-End heuristic and the DJ, DT and GEPETO clustering algorithms) were implemented within GEPETO and applied to the taxi dataset for identifying POIs. We used both the original and sanitized versions of the taxi dataset to evaluate the resilience of the inference attacks against sanitization. More precisely, we applied both sampling and perturbation techniques (cf. Section 3) with various ranges of parameters. In each situation, we evaluate the recall and precision of the produced POIs.

This recall-precision evaluation requires to be able to judge whether or not a POI is "correct". To automatize this process, we defined 6 areas in San Francisco that make good candidates for real POIs and, which are at the same time generic enough: the taxi company parking lot, the main train station, the airport, the city center and three entertainment areas (the Castro district, Fisherman's Wharf and the Golden Gate recreational park). The *precision* is defined as the ratio between the number of correct POIs and the total number of POIs returned by an algorithm. In our experiments, a POI is considered "correct" if it falls inside one of the 6 ground truth areas. The *recall* is the ratio between the number of area detected (i.e. hit by at least one POI) and the total number of areas. According to these definitions, an algorithm randomly generating many POIs would have a high recall and a low precision, as it would probably identify all the areas but many POIs would fall outside many of them. An "ideal" algorithm, displaying a high recall ad high precision, would generate 6 POIs, one for each area.

Figure 2 measures the recall-precision trade-off of the 4 algorithms against a sampling technique. We also evaluated "Natural", a naïve algorithm that directly outputs all the points of the dataset as POIs, which results in a low precision but a high recall. In Figure 3, the evaluation is performed for the 4 algorithms against random perturbation. These experiments have shown that the Begin-End heuristic has an excellent recall, due to the high number of POIs generated and an average precision but is sensitive to sampling. Indeed, when the sampling rate reaches the size of the time window of the begin-end heuristic, it considers all the traces as POIs and is as precise as the "Natural" algorithm. On the other hand, the begin-end heuristic is not too impacted by perturbation and even under large distortion, it stays one of the more precise algorithm. DJ Cluster displays a terrible behaviour in the presence of sampling, and even a worst one with respect to distortion. Indeed, the first phase of the algorithm removes the moving traces to focus on those where the individual is not moving. With sampling, the probability is high that static traces are removed by the sanitization process. Moreover, under the action of perturbation, every single trace implies some movement. Henceforth, all the traces are removed during the first phase of the algorithm and DJ Cluster does not output any POI. DT Cluster is highly resilient against sampling, with a high re-

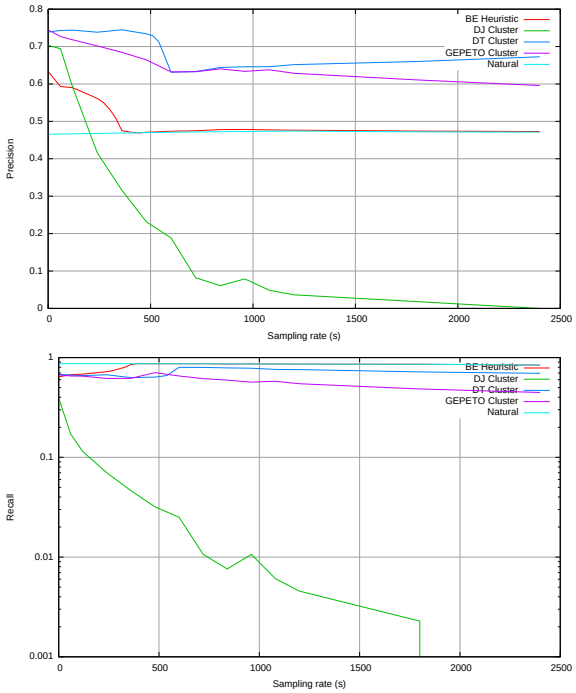

Figure 2: Precision-recall with sampling.

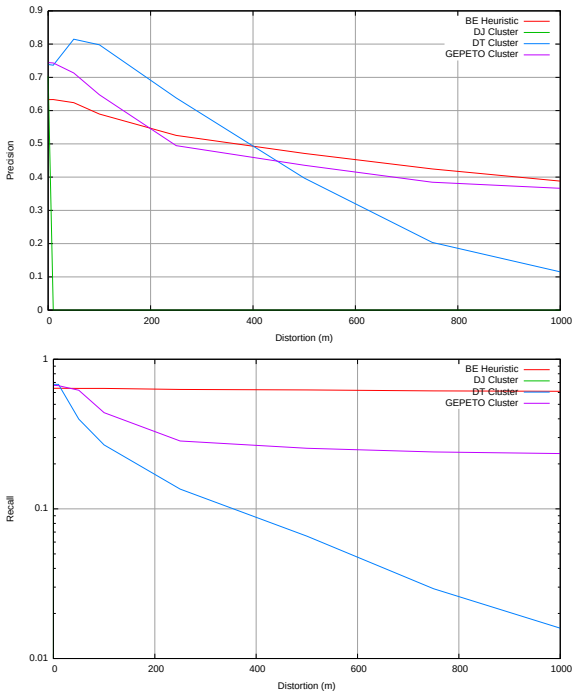

Figure 3: Precision-recall with perturbation.

call and the best precision, but displays a bad recall against distortion. However, the precision of the remaining POIs remains good under moderate distortion. Finally, GEPETO Cluster seems to be a good compromise. For instance, its behaviour is comparable or just below DT Cluster in the presence of sampling with average to good recall and precision. Moreover, under distortion, it seconds the Begin-end heuristic with an average recall and a high precision.

To summarize, the efficiency of the inference attack depends strongly on the sanitization process that has been performed on the target data. For instance, in the presence of sampling, the DT cluster algorithm offers a high recall and a good precision, but its performance degrades significantly with respect to distortion. Therefore, if the adversary knows *a priori* that sanitization is only based on sampling, then he can directly choose the DT cluster algorithm for performing an efficient inference attack. On the other hand, GEPETO cluster seems to be a reasonable alternative for an adversary having no knowledge of the sanitization technique applied as its performance remains good under both sampling and distortion.

5. CONCLUSION

From the point of view of the attacker, these experiments show that the behaviour of the clustering algorithms can diverge significantly depending of the circumstances, for instance when sanitization is applied. On the other hand from the point of view of the data curator looking for the best sanitization method to protect privacy while preserving some utility in the geolocated datasets published, the conclusion is different. For instance, both the DT and GEPETO clustering algorithms are quite resilient to sampling. Moreover, regarding perturbation, it seems that no clustering algorithm (among those we evaluate) performs a better precision than 50% under a distortion of magnitude 400 meters. A fundamental interrogation is whether or not the data is remains useful with such a high level of distortion. As proof of concept, these experiments have demonstrated the usefulness of GEPETO as a tool to evaluate various algorithms for attacking or sanitizing geolocated data, but of course this is only a first step and more exhaustive experiments, with more sophisticated inference attacks and sanitization methods, remain to be done. Moreover as briefly highlighted previously, there is a strong interplay between the geolocated data of an individual and its social network in the sense that knowledge about one can help infer new information about the other (and *vice-versa*). We plan to investigate the inference attacks combining location and social knowledge and integrate them in GEPETO.

Being able to *quantify privacy* with respect to a particular geolocated dataset is another fundamental issue as it can be used for instance to measure the privacy gained by using protection mechanisms (such as sanitization algorithms). Despite several propositions that can be found in the literature, the problem of finding relevant privacy metrics for geolocated data is still open for now. For instance, is an individual hidden inside a crowd gathered in a small area really more protected in terms of privacy than an individual alone in the middle of a large area such as a desert? Or should we rather define privacy according to how much the behavior of an individual is indistinguishable of the behaviors of other (or a group of) users? One possibility is to study how anonymity is defined in anonymous communication (e.g. the notion of anonymity set) [14] and how this applies to geolocated data. Taking into account unlinkability in the privacy metric seems to be particularly crucial in this context. Indeed, if someone can gather and link the movements of an individual during some period, he or she can build a complete profile of his behaviour if combined

with other inference attacks.

Due to lack of space, we have mainly focused on describing how to protect geoprivacy with sanitization procedures in this section but of course other approaches are also possible. For instance by using *cryptographic primitives*, ubiquitous systems can perform computations which depend on their geolocated data in a secure manner such that only the output of the global computation is learnt (and nothing else). Moreover, *access-control mechanisms* can be used to control how an external entity accesses the geolocated data of individuals within a system. By auditing queries, it also can decide whether or not it should disclose more information since this could cause a privacy breach. In some sense, these two approaches are complementary to the sanitization one.

6. REFERENCES

[1] S. Gambs, M.-O. Killijian, and M. N. del Prado, "GEPETO: a GEoPrivacy Enhancing Toolkit," in *Proceedings of the International Workshop on Advances in Mobile Computing and Applications: Security, Privacy and Trust, held in conjunction with the 24th IEEE AINA conference, Perth, Australia*, April 2010.

[2] J. H. Kang, B. Stewarta, G. Borriello, and W. Welbourne, "Extracting places from traces of locations," in *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, 2004, pp. 110–118.

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[4] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, June 2008.

[5] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[6] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," *Pervasive Computing*, pp. 390–397, May 2009.

[7] L. Jedrzejczyk, B. A. Price, A. K. Bandara, and B. Nuseibeh, "I know what you did last summer: risks of location data leakage in mobile and social computing," *Department of Computing Faculty of Mathematics, Computing and Technology The Open University*, November 2009.

[8] D. Ashbrook and T. Starner, "Learning significant locations and predicting user movement with *GPS*," in *Proceedings of the 6th IEEE International Symposium on Wearable Computers*, 2002, pp. 101–109.

[9] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Enhancing security and privacy in traffic-monitoring systems," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 38–46, 2006.

[10] Z. Changqing, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personal gazetteers: an interactive clustering approach," in *Proceedings of the ACM International Workshop on Geographic Information Systems*, 2004, pp. 266–273.

[11] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition using relational Markov networks," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005, pp. 773–778.

[12] D. Reiter, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.

[13] J. Krumm, "Inference attacks on location tracks," *Pervasive Computing*, pp. 127–143, 2007.

[14] A. Pfitzmann and M. Hansen, "Anonymity, unlinkability, unobservability, pseudonymity, and identity management – a consolidated proposal for terminology," February 2008.

[15] M. P. Armstrong, G. Rushton, and D. L. Zimmerman, "Geographically masking health data to preserve confidentiality," *Statistics in Medicine*, vol. 18, pp. 497–525, 1999.

[16] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," *Proceedings of the ACM/USENIX International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2003.

[17] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Computing*, pp. 46–55, 2003.

[18] T.-H. You, W.-C. Peng, and W.-C. Lee, "Protecting moving trajectories with dummies," in *Proceedings of the 2007 International Conference on Mobile Data Management*. IEEE Computer Society, 2007, pp. 278–282.

[19] F. Giannotti and D. Pedreschi, *Mobility, Data Mining and Privacy Geographic Knowledge Discovery*, 2008.

[20] D. Matt, K. Lars, and B. Athol, "A spatiotemporal model of obfuscation strategies and counter strategies for location privacy," *Lectures Notes in Computer Science*, vol. 4197, no. 4, pp. 47–64, 2006.

[21] B. A. Price, K. Adam, and B. Nuseibeh, "Keeping ubiquitous computing to yourself: A practical model for user control of privacy," *Int. J. Human-Computer Studies*, pp. 228–253, 2005.

[22] R. Hariharan and K. Toyama, "Project lachesis: Parsing and modeling location histories," *Lecture notes in computer science - Geographic information science*, vol. 3, pp. 106–124, October 2004.