Protecting privacy of GPS trails by anonymization of the road graph

Mariusz Kamola
Institute of Control and Computation Engineering
Warsaw University of Technology
00665 Warsaw, Poland
M.Kamola@ia.pw.edu.pl

ABSTRACT

The problem of anonymization of car GPS traces is addressed. Unlike the existing approaches, we propose to preserve anonymity of the whole road graph, while performing accurate GPS location projection onto that graph edges. Projected locations are denoted by their distances from graph vertices, which allows to perform many analytical tasks, in a graph abstracted from physical node locations. With adequately selected subset of samples, the released data will still be enough to perform refined analysis, e.g. of the driving style, while preserving the anonymity of the graph. Possible deanonymization attacks and countermeasures are discussed shortly.

CCS Concepts

•**Security and privacy** → **Pseudonymity, anonymity and untraceability;** •**Computing methodologies** → *Model development and analysis;* •**Social and professional topics** → User characteristics;

Keywords

GPS privacy, driving style, traffic measurement

1. INTRODUCTION

Proliferation of GPS receivers that report their location is a blessing for decision makers but, equally, it raises many privacy issues. More and more often, such location tracking can be done without user will, consent or even knowledge. Collecting traces is a fact, and firms that are in possession of such tracks are well aware of their both value and delicacy.

That is why so much research effort (cf. Sec. 2) has been spent on proper anonymization of user location, yet preserving it usable for analysis and decision making. The analysis may be focused on traffic topology (e.g. road planning), on individual behavior (insurance ranking), on social interactions (emergent behavior of the crowd) and so on. In fact, an anonymization approach should be adequate to the kind of planned analytic work on the dataset.

This paper presents an alternative car trace anonymization approach, while preserving most of the important data:

- location readings stay grouped in *trips*, i.e. sequences left by a given car (allows tracking driver behavior);
- trip start and end location stay revealed (allows grouping trips for social analysis);
- only readings containing relevant information about the driving style are revealed (allows analysis of driving safety and ecological impact of individual drivers).

The rest of the paper is organized as follows. Sec. 2 covers similar approaches to location anonymization problem. Sec. 3 contains the proposition of the anonymization approach, and Sec. 4 — analysis of real traffic data that supports the proposition. Sec. 5 concludes the work.

2. RELATED WORK

The topic of location anonymization has been under research for at least 10 years, with the main approach to obliterate accurate location whenever there are too few other traces in neighborhood. In other words, it is believed an individual can effectively "get lost in the crowd" — provided there is enough crowd around. This "enough" get expressed as k-anonymity, i.e. there should be at least k sufficiently similar individuals around to let the true location be revealed without obfuscation. An adequate anonymity measure, as well as the research context, is presented in [6], where also a concept of erasing some track data is introduced, such that an adversary follower gets lost. Another paper [9] underlines users' lack of concern for their privacy; it also classifies typical anonymization operations:

- deleting — removing track data for vulnerable regions (work/home neighborhood),
- randomizing — adding noise to measurements, wherever required,
- discretizing (cloaking) — aligning location to predefined grid points,
- subsampling — removing track data periodically,
- mixing — interchanging data between tracks (real or artificially generated ones).

© 2015 ACM ISBN 978-1-4503-3973-5/15/11...$15.00
DOI: http://dx.doi.org/10.1145/2835022.2835033

Many papers present new or existing anonymization approaches along with adequate deanonymization methods; most of the latter ones are based on some prior assumptions about user characteristics: driving style, points of interest (POI) visited, home and work location. With that extra data one may carry out a match which maximizes a probabilistic measure of similarity between track and the known user preferences, as in [5], where Markov chains represent users' mobility between POIs, thus almost uniquely defining their individual preferences. That is why some obfuscation approaches [3] apply sophisticated apparatus to map current location onto a different place in road graph, that provides effective anonymization but influences the quality of results as little as possible.

Instead of distorting the current location, one may effectively play with the timestamps of the existing samples, or replace some of the samples altogether with their synthetic counterparts, calculated using e.g. some shortest path routing algorithm [8]. Irrespective of the anonymization attempts mentioned here, one must be aware that the original k-anonymity measure refers to *similar* tracks in the neighborhood; otherwise, different driving style (e.g. for trucks, autos and motors) can easily betray the followed individual [10].

3. PROPOSED APPROACH

The common and indisputable assumption made so far is that the GPS traces must be anonymized within unchanged system of coordinates: latitude vs. longitude. We claim here that for many analytic tasks this is unnecessary, and that an alternative approach may be useful as well — where the position is expressed in terms of relative or absolute distance from nearest road junction or intersection. More formally, we define the mapping $\mathcal{M}_{G(V,E)}(p_1^\star, ..., p_K^\star) \to (p_1, ..., p_N)$ of a trip onto a road graph G. A trip is a sequence of recorded GPS locations $p_i^\star = (p_{\text{lat},i}, p_{\text{lon},i})$ — the trip index is omitted here. The output is also a sequence of location points, however, containing one-dimensional points positions within a specific graph edge, $p_j = (p_{\text{rel},j}, p_{\text{edge},j})$, where $p_{\text{edge},j} \in E$.

Not all original GPS locations have their mapped counterparts ($K \geq N$) but those who have get projected onto the graph as precisely as possible. The point is that the data owner is not going to reveal any particular information about the geographical location of road graph nodes to the organization performing data analysis. Although that organization is presented with trip data referring to an anonymous graph, it is still capable of performing many useful analytical tasks, because:

1. Actual route taken in a trip remains unaltered (in the topological sense). This allows to do may useful traffic analyses: determine busy intersections, model user maneuvers on intersections, and even determine neighborhoods (in terms of edge hops) for location of new investments.

2. Trip start, end and via points remain unaltered. This allows to group trips and infer about drivers' lifestyle.

3. Data best describing driving habits remain unaltered. This allows to reason about driving safety, unsocial behavior or ecological impact per single trip.

The first property is rather straightforward: note that limiting trip information in the released dataset to the ID of traversed edges, with no information whatsoever about edge length, makes it very difficult for the adversary to perform a match to known road graphs. Formally, it is the problem of finding a subgraph of a given topology, which is known to be a NP-complete task in general. Although linear-time matching algorithms exist [4], they are developed for planar graphs, and with limitation on the subgraph size. Their applicability for our problem still has to be verified. One should be also aware that traffic intensity on certain edges may give topological hints and make the deanonymization easier. At the same time, roads with no GPS traces on them may not appear in the released graph altogether, increasing the matching difficulty by a factor which apparently has not been yet studied.

The second property can easily be attained by revealing start, finish and stop points, and setting p_{rel} to the fraction of the edge length (road segment length) where the event took place. Certainly, data distribution in POI proximity helps realizing the segment length and make deanonymization easier. In such case, location alignment to car park center may be the countermeasure.

We consider the third property the hardest to attain because it is difficult to state beforehand which data are most representative for driving style. Releasing too many of them would greatly help deanonymization, giving hints about road segment (graph edge) length. We have therefore studied a real case to find that out.

4. IMPLEMENTATION

The data being subject to analysis contained raw GPS traces of trips made in two days in a Polish city and its suburbs (20 by 20 km). Car locations were collected every 10 sec. and presented natural inaccuracies and data loss. The initial dataset contained over 3 million GPS readings, and over 19,000 trips. Projection of GPS location on the Openstreetmap (OSM) road graph was aided by GraphHopper (GH) library [1], which associates locations with best matching road graph edge. That map matching procedure is not a trivial one as it involves recursive shortest path planning for subsets of the data, leaving out readings that would definitely spoil a common-sense route (this excludes some scenarios involving a detour e.g. dropping someone at the airport "kiss and fly" section). Eventually, 13,000 trips were processed correctly, using 1/3 of the original amount of data.

Since GH does not do actual point projections, we have done this in postprocessing phase. Road graph edge is represented as a sequence of *pillars* (intermediate points) marking exact trail between *towers* (junctions or intersections); we have projected GPS location orthogonally on the resulting sections or pillars/towers — whichever was closest. Each original GPS reading contained also instantaneous speed, calculated internally by the device; we preserved that data for further analysis.

In order to conceal edge length while exposing representative samples, we have to investigate, which samples contain most information about the driving style. We propose to divide the process of passing a graph edge (road segment) into three phases: initial, intermediate and final. Our hypothesis is that driving in the intermediate phase is more stable than in the two other phases, carries few information and can safely be left out. At the cost of some data loss, we could release data for the initial and final phases

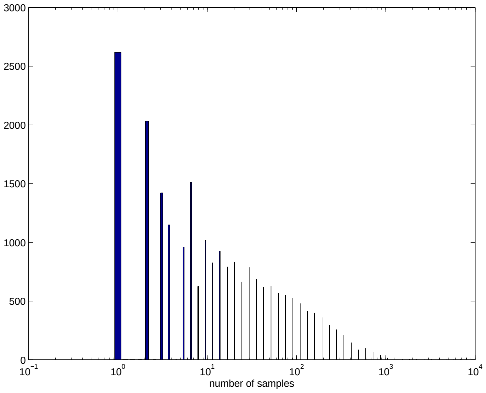

Figure 1: Histogram of the number of samples per road graph edge.

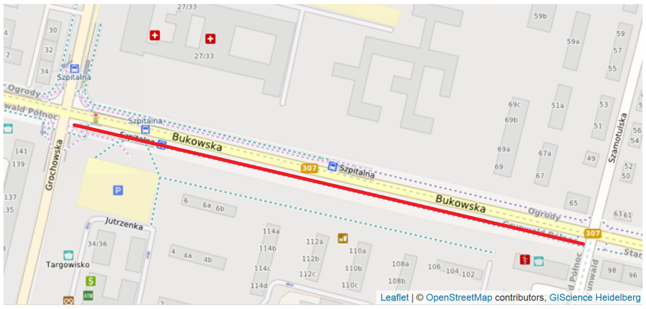

Figure 2: Screenshot of the street section in question (marked with thick red line). Adjacent intersections have traffic lights; no stopping is allowed for the whole section.

only, thus depriving the adversary of the information about the edge length. To verify that, let us consider only data from edges having enough traffic, to reduce sampling error. In the histogram of samples number per edge (Fig. 1), we see the linlog decrease: taking 30 samples per edge for the threshold leaves us still pretty 88% of the original data to work with. Next thing is the size of the initial and intermediate phases. Rough estimates of fast car acceleration process (2.5m/s^2 and typical speed limit 50 km/h) give us 60 meters as the absolute minimum. Then observation of empirical speed probability distribution functions (Fig. 4) for various phase lengths turns out to be a very informative one. The samples were taken for an edge of 460 m, between two intersections with traffic lights, with one lane in each direction (Fig. 2). Interestingly, the initial phase is a very uniform one: almost all vehicles attain the speed of 40 km/h. In the intermediate phase, the speed grows a little, but also it disperses in both directions (due to overtaking and bypassing buses stopping in dedicated bays). Finally, we observe many full stops in front of the other lights; but also quite a few vehicles passing fast. Formally, for all the three phase lengths considered, the final phase speed distribution is most dispersed, supporting the hypothesis that it may convey enough information about individual style of driving. On the contrary, the initial phase gives much less opportunity for certain drivers to "express themselves". Certainly, this may differ for various types of road designs, and the intersections with many lanes and light traffic should get more attention.

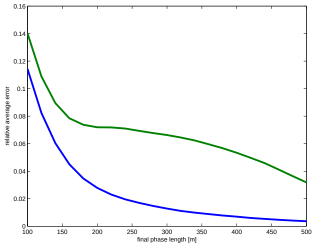

Figure 3: Average errors for various length of the final phase: maximum velocity error (lower, blue) and maximum speed variability (upper, green).

We propose two metrices of driving dynamics, and calculate estimation error in case of the proposed final-phase-only approximation. One is just the maximum velocity reached while traversing a graph edge. In unchanging conditions, drivers differ in achieved speed, due to both intrinsic and extrinsic conditions [10]. Another is the maximum acceleration or deceleration, measured as difference of instantaneous speed for two consecutive samples — also calculated for a graph edge. Rapid braking in final phase may indicate bang-bang style of velocity control, with the risk of passing at red light, with travel time as the only goal. In Fig. 3 we present relative mean errors for both metrices, for various lengths of the final phase. The numbers were calculated only for those edge traversals with three or more samples available, where the last two belong to the final phase. Setting the final phase length to 150 m gives substantial reduction for both errors, which can still be reduced but at unproportionally larger cost.

5. CONCLUSIONS

The main contribution of this paper is the postulate that many traffic analyses can be carried out for an abstract road graph, thus letting the provider release exact data while preserving privacy. The abstract graph is deprived of GPS coordinates for nodes. Deanonymization via topological match to the much larger road graph is prohibitively complex. If the released data contain only GPS samples projected for the final 150 m of an edge , then that data, along with the momentary speed, carry much information about the particular driving style.

There are still open issues related to security of such approach. Short edges have not enough samples and can be easily identified — we propose to i) delete them altogether or ii) merge them with adjacent ones if there is no traffic split at the junction or iii) synthesize enough samples using some analogy or learned probability model of the driver. Long edges with heavy traffic are likely to be recognized by a human expert — we propose to split them artificially, or to add some faked subgraph — but, consequently, such distortion should be done by a human. For some very specific urban layouts (dense grids connected by few links, cf. eg. New York City topology, or coastal locations, in general) the proposed approach probably will not give satisfactory results.

Another possible attack scenario is to look for known trips in the graph, and thus deanonymize the initial set of edges. It is, again, subgraph search problem, with the subgraph

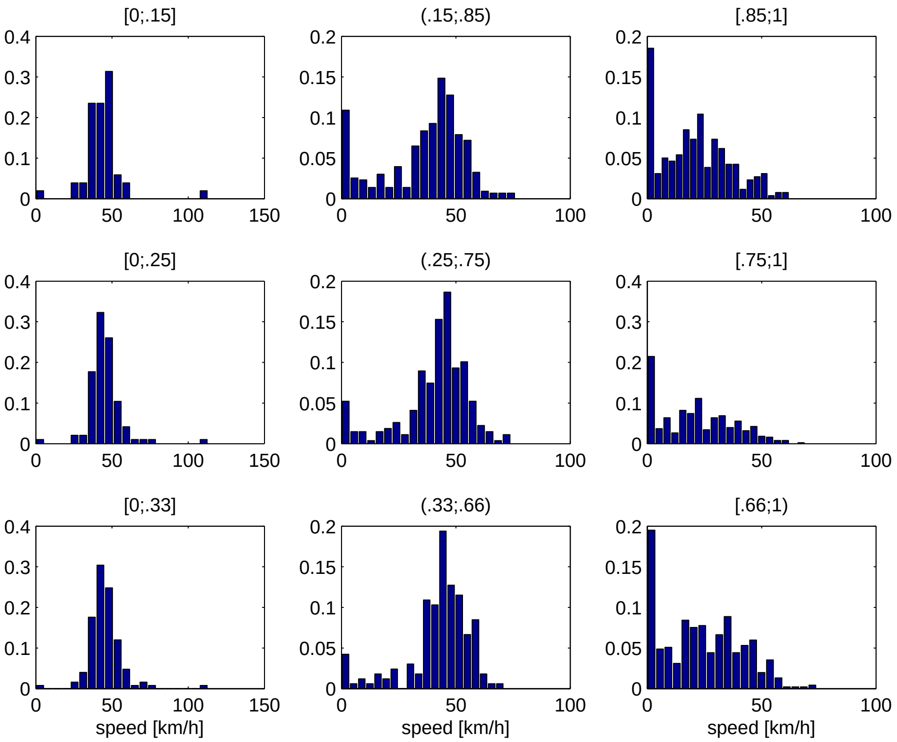

Figure 4: Probability distribution functions for the selected road section of 460 m. Graph titles denote intervals (expressed as fractions of the section length). Experiments were carried out for initial, intermediate and final phases (columnwise), for various phase lengths (row-wise; the middle interval increasing).

being the known trip. If it has a unique shape, it will compromise the road graph effectively. This is equivalent to Sybil attack in social network [2]. The solution at hand is to detect unique trips, and purge them from the dataset, so that k-anonymity be preserved w.r.t. trip topology.

It should be reminded that our approach is intended to be used in situation when analysis is outsourced to some organization, which either will not have access to the data when true locations will be made known to the final customer, or will never know the identity of that final customer. So, the more probable implementation is to make that organization work only on the dedicated infrastructure of the data provider — cf. the physical layout of the SUMA project [7].

6. ACKNOWLEDGEMENTS

This work was partially funded by the Polish National Center for Research and Development under grant agreement DOBR/0071/R/ID 1/2012/03.

7. REFERENCES

[1] Graphhopper homepage, www.graphhopper.com.

[2] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou R3579X?: Anonymized social networks, hidden patterns, and structural steganography. *Commun. ACM*, 54(12):133–141, Dec. 2011.

[3] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Optimal geo-indistinguishable mechanisms for location privacy. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 251–262. ACM, 2014.

[4] D. Eppstein. Subgraph isomorphism in planar graphs and related problems. *CoRR*, cs.DS/9911003, 1999.

[5] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614, 2014.

[6] B. Hoh and M. Gruteser. Preserving privacy in gps traces via uncertainty-aware path cloaking. In *In Proceedings of ACM CCS 2007*, 2007.

[7] A. Kozakiewicz and K. Lasota. Secure DRM mechanism for offline applications. In *Proceedings of the International Conference on Military Communications and Information Systems ICMCIS (former MCC)*, 2015.

[8] V. Primault, S. B. Mokhtar, C. Lauradoux, and L. Brunie. Time distortion anonymization for the publication of mobility data with high utility. *arXiv preprint arXiv:1507.00443*, 2015.

[9] S. Wilson, J. Cranshaw, N. Sadeh, A. Acquisti, L. F. Cranor, J. Springfield, S. Y. Jeong, and A. Balasubramanian. Privacy manipulation and acclimation in a location sharing application. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 549–558. ACM, 2013.

[10] B. Zan, Z. Sun, M. Gruteser, and X. Ban. Linking anonymous location traces through driving characteristics. In *Proceedings of the third ACM conference on Data and application security and privacy*, pages 293–300. ACM, 2013.