

# Patterns

## The risk of re-identification remains high even in country-scale location datasets

### Highlights

- Re-identification risk is statistically modeled and shown to decrease slowly with dataset size
- With increasing dataset size, the unicity decrease is lower-bounded and convex
- Previous estimates of unicity unrealistically underestimated the risks
- Individuals are likely re-identifiable in country-size location data and other high-dimensional datasets

### Authors

Ali Farzanehfar, Florimond Houssiau,  
Yves-Alexandre de Montjoye

### Correspondence

demontjoye@imperial.ac.uk

### In Brief

Researchers have claimed that individuals could not be re-identified in large-scale location datasets, making them safe. We here empirically measure and mathematically model the relationship between the size of a dataset and the risk of re-identification. Our results show that the risk decreases slowly with dataset size, making even large country-scale datasets very likely to be re-identifiable.



## Article

# The risk of re-identification remains high even in country-scale location datasets

 Ali Farzanehfar,<sup>1</sup> Florimond Houssiau,<sup>1</sup> and Yves-Alexandre de Montjoye<sup>1,2,\*</sup>
<sup>1</sup>Department of Computing, Imperial College London, London SW7 2AZ, UK

<sup>2</sup>Lead contact

 \*Correspondence: [demontjoye@imperial.ac.uk](mailto:demontjoye@imperial.ac.uk)
<https://doi.org/10.1016/j.patter.2021.100204>

**THE BIGGER PICTURE** Data about us are being collected in many different ways, when we use our bank cards, use our phones, browse the web, or even drive our cars. These datasets contain detailed information about our lives. For each person, a dataset might contain thousands to tens of thousands of records. Previous research has shown that knowing just a few points about a target can single out the vast majority of people in location datasets. However, some had argued the risk of re-identification becomes negligible if we look at large-scale datasets containing tens of millions of people.

Here, we empirically measure, mathematically model, and provide a lower bound on the relationship between the size of a dataset and the risk of re-identification. Our results all show that re-identification risk decreases very slowly with increasing dataset size. Contrary to previous claims, people are thus very likely to be re-identifiable even in country-scale datasets.



**Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

## SUMMARY

Although anonymous data are not considered personal data, recent research has shown how individuals can often be re-identified. Scholars have argued that previous findings apply only to small-scale datasets and that privacy is preserved in large-scale datasets. Using 3 months of location data, we (1) show the risk of re-identification to decrease slowly with dataset size, (2) approximate this decrease with a simple model taking into account three population-wide marginal distributions, and (3) prove that unicity is convex and obtain a linear lower bound. Our estimates show that 93% of people would be uniquely identified in a dataset of 60M people using four points of auxiliary information, with a lower bound at 22%. This lower bound increases to 87% when five points are available. Taken together, our results show how the privacy of individuals is very unlikely to be preserved even in country-scale location datasets.

## INTRODUCTION

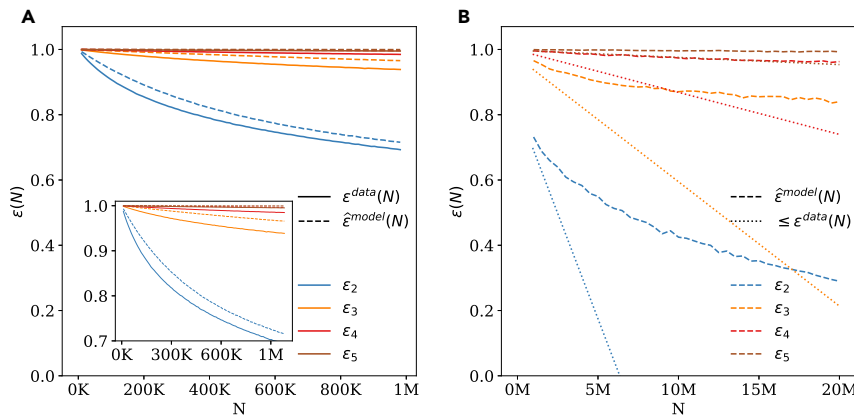
Throughout our day, we interact with many digital services when using our phone, paying with our credit card, or using public transport with a smart card. This results in our location data being collected broadly, sometimes on the scale of countries. For instance, Vodafone UK collects location trajectories of 20M citizens<sup>1</sup>—a third of the population—while up to 5 million people use London's subway daily.<sup>2</sup>

Location data have been used extensively in research. In urban planning, mobility data can be used to monitor urban activity<sup>3</sup> and help design better cities.<sup>4</sup> In epidemiology, it has been used to monitor and mitigate the spread of infectious diseases such as

Ebola and COVID-19.<sup>5–10</sup> In computational social science, it has allowed us to gain unprecedented insights into the spatial distribution of poverty,<sup>11</sup> and even to study the impact of mass employment layoffs on society.<sup>12</sup> Further, the use of location data has withstood scrutiny into potential biases in their collection mechanisms.<sup>13</sup>

Despite this, the large-scale collection and use of location data has raised serious privacy concerns. It consists of fine-grained records of where we are and how we move around, and was considered sensitive by 82% of Americans in a recent survey.<sup>14</sup> Location data can furthermore be used to predict individuals' income,<sup>11,15</sup> their home and work locations,<sup>16–21</sup> when they sleep and wake up,<sup>22–26</sup> their gender and age,<sup>27</sup> their personality,<sup>28</sup> who their friends are,<sup>29,30</sup> and where they tend to socialize.<sup>31</sup>





**Figure 1. The relationship between unicity and dataset size**

(A) Empirical (solid lines) and estimated (dashed lines) unicity decreases slowly with the size of the dataset. Inset: close up of the region  $\epsilon \geq 0.7$ .

(B) The estimated unicity remains high even in large datasets. This is confirmed by the lower bound results (dotted lines). Taken together, these results strongly suggest that unicity remains high even in country-scale datasets.

## RESULTS

Our experiments are performed on a dataset of call detail records containing the location of 1M individuals over 3 months. Each record contains a unique user ID,

unicity has been proposed as a measure for the risk of re-identification in anonymous datasets and was used to show how four points of auxiliary information (places and times where someone was) are enough to uniquely identify 95% of people in a large-scale location dataset.<sup>32</sup> These four points of auxiliary information could be in the form of geo-tagged “tweets,” online check-ins, or information obtained by more traditional means, such as observing someone making a call. Unicity ( $\epsilon_p$ ) is defined as the fraction of trajectories that are unique based on knowledge of  $p$  randomly chosen points in a given trajectory. Unicity has since been used to quantify re-identification risk across a number of domains, including the mobility of vehicles,<sup>33</sup> apps downloaded by smartphones over time,<sup>34,35</sup> smart cards used in public transport,<sup>24</sup> credit card transaction histories,<sup>36</sup> and location data from mobile phones in a number of countries.<sup>32,37,38</sup> A range of studies have furthermore exploited the unicity of datasets to re-identify people. Narayanan and Shmatikov famously showed that close to 90% of people could be re-identified in the Netflix dataset,<sup>39</sup> while Riederer and colleagues used the unicity of traces to match the same individual across multiple datasets.<sup>40</sup>

Researchers and industry practitioners have, however, argued that these high unicity numbers are an artifact of the small size of the datasets considered, and are overestimating the risk of re-identification.<sup>41–43</sup> For instance, Riederer et al.<sup>40</sup> relied on a location dataset of 1.7k people, while other case studies report unicity on dataset sizes ranging from several thousands (respectively 12k and 55k)<sup>33,34</sup> to over 1 million people (1.5M).<sup>32</sup> Examining a published study,<sup>36</sup> El Emam et al. estimated that the unicity of a dataset of  $\approx 20M$  trajectories will be as low as 1% given four points of auxiliary information, the conclusion being that privacy was preserved in such large datasets.<sup>42</sup>

We here (1) study 3 months of location data and show empirically that unicity decreases slowly with the size of the dataset, (2) approximate this decrease with a simple statistical model taking into account three population-wide marginal distributions along with the underlying geography, and (3) prove that the decrease in unicity is a convex function of the dataset size and obtain a linear lower bound on unicity. We finally perform a sensitivity analysis suggesting that the decrease in unicity is agnostic to broad perturbations in the input distributions. These results disprove previous claims, instead showing that unicity is likely to remain high even in country-scale datasets.

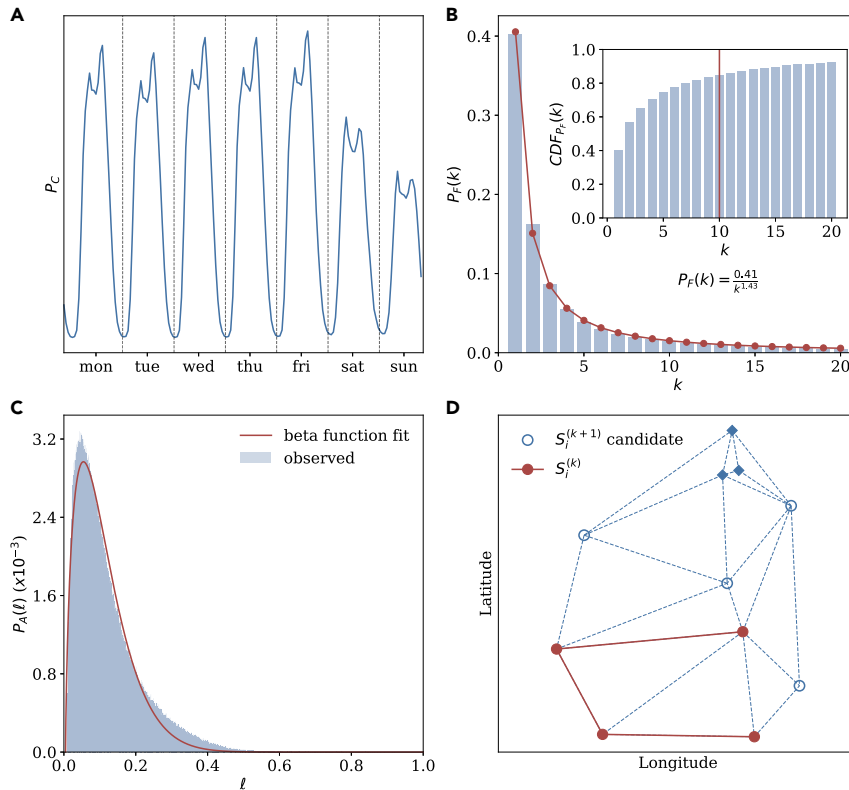
an hourly time stamp, and an antenna ID, which relates to a location (see [Supplemental information](#) for more details). We formally model this dataset as a sequence,  $D = (D_1, \dots, D_N)$ , populated with user time/location traces of the type  $D_i = (X_i, C_i)$ .  $X_i$  and  $C_i$  are lists of positions (antennas) and times (hours) representing the spatial and temporal components of a user’s location trace.

Using this dataset, we empirically study the decrease in unicity with the dataset size by randomly sampling individuals from our original dataset and measuring the unicity of the sample as we increase its size (see [Experimental procedures](#) for details). We use the formal definition of unicity and the estimation algorithm S2 from de Montjoye et al.<sup>36</sup> In line with previous work, we use the subscript  $p$  in  $\epsilon_p(N)$  to indicate the number of points of auxiliary information used in the computation of unicity.

Figure 1A shows that unicity empirically decreases slowly with the size of the dataset. With three points of auxiliary information, unicity (solid orange line) goes down from  $\epsilon_3(100K) = 0.98$  in a dataset of 100,000 people to  $\epsilon_3(1M) = 0.93$  in a dataset of a million people. With two points (solid blue line) this decreases slightly faster, reaching  $\epsilon_2(1M) = 0.69$ , while unicity with four points or more (solid red and brown lines) decreases very slowly with  $\epsilon_4(1M) = 0.98$ . These results show that, while the size of the dataset has an impact on unicity, the decrease in unicity is slow.

To further study how unicity decreases with dataset size and whether it decreases sufficiently in population-scale datasets, we propose a simple statistical model taking into account three population-wide marginal distributions—circadian ( $P_C$ ), frequency ( $P_F$ ), and activity ( $P_A$ )—along with the network of mobile phone antennas in a country. Using solely these quantities, the model is able to replicate the observed decrease in unicity with dataset size.

Figure 2 displays the information extracted from the dataset, three distributions, and the antenna network. ( $P_C$ ) characterizes the circadian cycle, the overall likelihood of a record to occur at a given time in a week. The existence of circadian cycles is well documented in the computational social science literature,<sup>22,23,25,26</sup> and we use their empirical form in the model. The frequency distribution, ( $P_F$ ), is the relative overall likelihood of a location to be visited. This distribution too has been studied before and has been widely shown to be well approximated by a power-law distribution,<sup>44–48</sup> as is also the case here (Figure 2B,  $R^2 = 0.99$ ). The activity distribution, ( $P_A$ ), captures the number



**Figure 2. Inputs to the unicity model**

(A) The circadian distribution,  $P_C$ . (B) The frequency distribution,  $P_F$ , along with a power law fit (solid line,  $R^2 = 0.99$ ). The inset displays the cumulative distribution with 85% of activity captured by the top 10 locations. (C) The activity distribution,  $P_A$ , indicating the distribution of the number of records per trajectory along with a  $\beta$  distribution fit (solid line,  $R^2 = 0.98$ ). (D) Illustration of the sub-graph sampling method used to generate an antenna set  $S_i$  where  $S_i^{(k)} \in S_i$ . The underlying antenna network is represented by dotted lines. The filled nodes (circles) correspond to locations already selected, while the hollow nodes are potential locations that could be selected next ( $S_i^{(k+1)}$  candidates) (see [Supplemental information](#) for detailed algorithm). Remaining locations are represented by filled diamonds.

be expressed as a sum of convex functions of  $N$ , and is thus convex.

This builds on two assumptions: (1) there exists an underlying trajectory distribution  $T_X$  from which all trajectories  $D_i \in D$  are sampled and (2) all trajectories are independent of one another,  $D_i \perp D_j$ . The first assumption states that an underlying distribution for trajectories exists. Such a distribution would also capture correlations between individuals on a large scale (e.g., commuting patterns, cities, weekends). The second assumption presumes that the correlation between specific individuals is negligible when estimating unicity of large datasets.

A direct consequence of unicity being a strictly decreasing convex function is that it will be lower bounded by its linear tangent (treating unicity as a function of a real-valued  $N$ ):

$$\epsilon(D(N)) \geq \epsilon(D(N')) + (N - N') \cdot \left. \frac{d\epsilon}{dN} \right|_{N=N'} \quad (\text{Equation 1})$$

Re-arranged and expressed for discrete values, this gives a lower bound for unicity:

$$\epsilon(D(N')) - \epsilon(D(N)) \leq (N - N') \cdot (\epsilon(D(N' - 1)) - \epsilon(D(N'))) \quad (\text{Equation 2})$$

Using the tangent to the empirical unicity curves estimated by discrete difference over the range of  $N \in [0.9M, 1M]$ , we obtain a lower bound of 0.73 for  $\epsilon_4(20M)$  and 0.9 for  $\epsilon_5(20M)$  (Figure 1B, dotted lines).

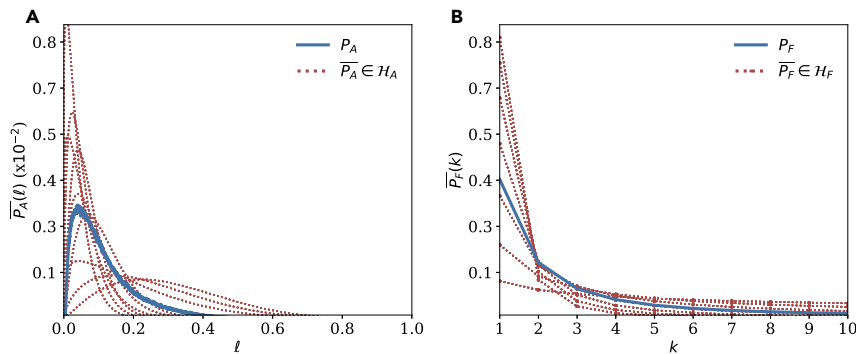
Our results show that unicity decreases slowly with the size of the dataset and that it, very likely, remains high even in population-scale datasets. This refutes previous claims that privacy is preserved in population-scale datasets, instead showing the risk of re-identification to be high. Modern location datasets have a great potential to improve our society, for example, by training AI algorithms, but robust privacy engineering solutions are needed to use them safely.

of records  $\ell_{(i)} = |D_i|$  that appear in each user trace. We approximate it here with a  $\beta$  distribution ( $\alpha = 1.72$ ,  $\beta = 14.7$ ,  $R^2 = 0.98$ ). Finally,  $S_i$  is the set of locations visited by person  $i$ . It is a sub-graph sampled from the Delaunay tessellation of the antenna coordinates ( $\mathcal{L}$ ) in the dataset (see [Supplemental information](#) for the detailed algorithm).

In short, for each user, our model samples a list of 10 connected antennas ( $S_1, \dots, S_{10}$ ) on the network and an activity (number of records in the user's trace),  $A \sim P_A$ . Each record's timestamp  $C$  and position  $X$  is then sampled according to the circadian distribution  $C \sim P_C$  and  $X = S_K$ ,  $K \sim P_F$ . This model is formally defined in the [Experimental procedures](#).

Figure 1A shows that our simple statistical model closely follows the empirical measure of unicity from 1 to 1M people (dashed and solid lines). Using the model, we then study how unicity is likely to evolve as the size of the dataset increases to 20M people (Figure 1B). For  $N = 20M$ , our model estimates unicity with three points to be close to  $\hat{\epsilon}_3(20M) = 0.93$ , while knowing one more point would increase this to the region of  $\hat{\epsilon}_4(20M) = 0.99$ . This is a stark difference with the linear extrapolation made by El Emam,<sup>42</sup> who reports a unicity of 0.01 with four points (we replicate El Emam's method in the discussion and display our results for up to 60M people in the [Supplemental information](#)).

The model provides good evidence that unicity is likely to remain high even in datasets as large as 20M people. For further evidence, we prove that the decrease in unicity with increasing dataset size follows a convex form, and use this result to provide a lower bound on unicity in large datasets. We show in the [Supplemental information](#) that the unicity of a dataset of size  $N$  can



**Figure 3. Range of distributions studied for the sensitivity analysis**

The ranges of perturbed activity  $\overline{P}_A$  (A) and frequency  $\overline{P}_F$  (B) distributions are displayed (dotted lines) along with their empirical forms (solid lines).

## DISCUSSION

Taken together, these results show that the scale of a dataset does not prevent re-identification. Human mobility, much like a physical fingerprint, is highly unique and can be used to find a person across mobility datasets.

Legally, the European Union (EU) General Data Protection Regulation sets a high threshold for what constitutes anonymous data, namely that the individual should not be identifiable taking into account both the “available technology at the time of the processing” but also future “technological developments” (Recital 26). The Article 29 Working Party, the predecessor to the European Data Protection Board, in its guidance sets out three criteria to assess whether a dataset is anonymous, singling out, linkability, and inference<sup>49</sup> with the former two being directly applicable here. As an example, the Centre for Humanitarian Data of the United Nations (UN OCHA) adopted 5% as a threshold for what constitutes an acceptable re-identification risk.<sup>50</sup> Even our lower bound of 22% far exceeds this liberal threshold.

Finally, here we study the unicity of location datasets with a spatial resolution of  $\approx 1 \text{ km}^2$  and a temporal resolution of an hour. Fine-grained GPS data are likely to lead to even higher values of unicity, and previous research has shown that, in general, de-identification methods do not meaningfully reduce the risk of re-identification. For instance, research<sup>32,34</sup> has shown that reducing the spatial and temporal resolution of the data further only slowly decreases the risk, while another study<sup>51</sup> concluded that location data “show poor anonymizability [as measured by  $k$ -anonymity], i.e., require important spatial and temporal generalization in order to slightly improve user privacy”.

Ensuring that these data can be accessed and used broadly is of paramount importance, but this should not come at the expense of people’s privacy. A range of privacy engineering techniques allowing data to be used while giving individuals strong privacy guarantees have been developed and are starting to be used.<sup>52–54</sup> As standards for anonymization are being redefined, in the EU and around the world, it is essential for them to emphasize the strong limits of de-identification, possibly banning the uncontrolled release of individual-level de-identified data, and to give guidance on the use of modern privacy-engineering solutions.

In the next three sections we discuss the underlying assumptions of the unicity model and some considerations regarding the sensitivity of our results and, finally, include a discussion on previous estimates of unicity.

## Assumptions underpinning the simple unicity model

We here evaluate the four assumptions underpinning the simple unicity model we present.

First, the model treats each of the four inputs in Figure 2 as independent of one another. Considering them, or some of them, jointly might further improve the model. This would, however, also increase its complexity and, therefore, its sensitivity to small changes in the data. Although further exploration would be interesting, we consider that the simple model approximates the decrease in unicity with increasing dataset size well enough to support our conclusion that unicity is unlikely to be low even in population-scale datasets.

Second, our model uses input distributions extracted from a dataset of 1M people to study the unicity of datasets with up to 60M people (see Supplemental information). This assumes that these distributions estimated from a smaller sample are representative of the larger sample (i.e., the estimation of the distributions has converged). We show that this is a reasonable assumption by instantiating our model  $\mathcal{M}$  with distributions extracted from samples of sizes significantly smaller than 1M, and showing that the unicity results remain largely unchanged (Figure S5 in Supplemental information). We also perform a sensitivity analysis to evaluate the impact of broad variations on these input distribution on our results (see next subsection).

Third, the model assumes each trajectory to contain at most one unique location. This allows for the mean frequency distribution ( $P_F$ ) to be used in the modeling process (Figure 2B). As seen in the inset of Figure 2B, more than 85% of the activity in the average trajectory is captured by the top 10 locations visited. Furthermore, we find that  $P_F$  changes only slightly when the number of unique locations is altered, and that our conclusions are not influenced by this choice.

Finally, our model assumes that the set of locations appearing in each trajectory can be described by a connected planar subgraph of the underlying antenna network. We believe this to be a reasonable assumption, as previous work suggests that subgraphs spanned by each trajectory in human mobility are highly localized, with the distribution  $P(r_g)$  of the radius of gyration—a metric for how far people tend to travel on average—following a power law with increasing radius.<sup>55</sup>

## Sensitivity analysis

Our simple statistical model for unicity takes as input three distributions. However, these distributions may vary depending on specifics of the dataset, such as the country where it was collected or the sources of location information. Here we perform a sensitivity analysis to ensure the robustness of our model to even broad changes to the distributions.

We first perturb the  $P_A$  and  $P_F$  distributions (Figure 3) around their empirical forms using a scaled earth mover’s distance as

**Table 1. Summary of unicity results at  $N = 20M$  as per the sensitivity analysis**

	$\epsilon_2$	$\epsilon_3$	$\epsilon_4$	$\epsilon_5$
Mean	0.307	0.735	0.876	0.935
Standard deviation	0.175	0.216	0.159	0.113
Minimum	0.071	0.260	0.431	0.544
Maximum	0.704	0.997	1	1

the guiding metric (see [Supplemental information](#) for details). The  $P_C$  distribution, on the other hand, has been shown to be very stable across datasets<sup>22–26</sup> and we thus keep it constant throughout our analysis.

These distributions are combined to produce 63 different instantiations of the unicity model ([Figure S2](#)). [Table 1](#) summarizes the unicity values for models using the broad range of distributions in [Figure 3](#), at a dataset size of 20M trajectories (see [Supplemental information](#) for 60M results). Note that the lowest unicity values across all instantiations of the model are still high, with  $\text{Min}(\epsilon_4(20M)) = 43.1\%$  and would still be considered as putting people’s privacy at risk.

Further, we study how certain aspects of human mobility contribute to unicity. Starting from empirical user location traces  $D_i = (X_i, C_i)$ , first, we find that removing the association between times ( $C_i$ ) and locations ( $X_i$ ), by shuffling the vectors and recombining them, only slightly affects unicity values ([Figure S4A](#)). Specifically, consider a dataset  $D'$  composed of trajectories  $D'_i = (X_i, C_i)$  such that:

$$X_i = \sigma_i(X_i),$$

$$C_i = \pi_i(C_i),$$

where  $\sigma_i$  and  $\pi_i$  refer to random permutations of the spatial and temporal components of  $D_i$ . This only marginally affects unicity, showing that unicity does not depend on the specific places being visited at specific times, as long as those times and places appear in the trace with their respective frequencies independently.

Second, we replace the set of locations in each trajectory with uniformly picked locations. Instead of using the sub-graph sampling method displayed in [Figure 2D](#), we populate each  $S_i$  with antennas picked from the entire set of locations  $\mathcal{L}$  uniformly at random. We find that this leads to unicity being overestimated ([Figure S4D](#)).

Third, replacing  $P_C$  or  $P_F$  with uniform distributions ([Figures S4B](#) and [S4C](#)) or attempts to model unicity using a simple combinatorial model ([Figure S3](#)) also cause the model to overestimate unicity. These demonstrate the importance of all three distributions and the underlying geography to correctly capture the unicity of mobility datasets.

This analysis, combined with the relative simplicity and generality of the unicity model, strongly suggest that our results would generalize to any location dataset. Likewise, the strong underlying combinatorial effect that underpins unicity combined with previous research<sup>34–36</sup> suggests that unicity will similarly decrease slowly in other types of high-dimensional data.

### El Emam’s method

El Emam<sup>42</sup> proposed a method (hereafter the EE method) to estimate the uniqueness of a population-size ( $N$ ) dataset given the unicity  $\epsilon(m)$  of a smaller sample dataset of size  $m$ . Using this method, he estimates that the uniqueness for a population of size  $N = 22 \cdot 10^6$  is about 1%, given a uniqueness of 90% of a sample of size  $m = 22 \cdot 10^6$  of the same dataset. This estimate forms the basis for his claim that uniqueness is low in large-scale datasets.

We here show that the EE method (1) is unrealistic and (2) provably gives the lowest possible estimate for the risk in the larger dataset, and that (3) by using our dataset, we observe that the real empirical unicity is significantly higher than the upper bound given by the EE method.

First, the method is unrealistic, as it effectively generates a dataset  $D$  of size  $N$  where a fraction  $\alpha$  of records are unique, while all the other records are identical to exactly one and only one other record. The parameter  $\alpha$  is selected such that the expected estimated uniqueness on a sample of size  $m$ , which we denote by  $\nu_D(m)$ , is equal to the empirical unicity. This assumes that users in the real mobility dataset are either unique or exact duplicates of another user.

Second, we prove in the [Supplemental information](#) that the risk estimated by the EE method will be lower or equal to the risk of *any* other dataset of size  $N$ , as this estimate is an *affine* function of  $m$ . In other terms, this method will *always* return the absolute lowest possible estimate of the risk.

Third, we apply the EE method to our dataset and show that its estimate of the risk is significantly lower than the real empirical value, leading to the risk of re-identification being strongly underestimated. For a dataset of 200,000 people, we empirically observe an  $\epsilon_2(200K) = 0.86$ . Using this number, El Emam’s method would estimate the risk of a larger 1M person dataset to be  $\epsilon_2(1M) = 0.3$ , while the correct empirical value is  $\approx 0.7$ .

Taken together, our results cast serious doubt on the validity of the EE method to carry out risk assessments.

## EXPERIMENTAL PROCEDURES

### Resource availability

#### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yves-Alexandre de Montjoye ([demontjoye@imperial.ac.uk](mailto:demontjoye@imperial.ac.uk)).

#### Materials availability

There are no physical materials associated with this study.

#### Data and code availability

Due to reasons of confidentiality and user privacy, we cannot share the raw data. However, we can make available all the input distributions and raw empirical results upon request for purposes of reproducibility.

The code used for all experiments is available at: [github.com/computationalprivacy/scaling-unicity](https://github.com/computationalprivacy/scaling-unicity).

### The unicity model in detail

We propose a simple statistical model  $\mathcal{M}$  taking into account three population-wide distributions: activity ( $P_A$ ), circadian ( $P_C$ ), and frequency ( $P_F$ ). This model samples location traces for each user independent of other users to estimate unicity of a dataset of size  $N$ . These location traces are then grouped together to compute unicity.

Formally, the model  $\mathcal{M}$  can be written as:

$$\mathcal{M}(P_A, P_C, P_F, \mathcal{L}, N) = D = (D_1, \dots, D_N). \quad (\text{Equation 3})$$

Each  $D_i \in D$  is a location trace for a unique user, represented as a list of  $L_i$  records  $(X_i^{(j)}, C_i^{(j)})_{j=1}^{L_i}$ . The length  $L_i$  of trace  $D_i$  is sampled from the empirical activity distribution  $P_A$ :

$$\mathbb{P}[L_i = \ell] = P_A(\ell). \quad (\text{Equation 4})$$

The timestamps of each record in a trace,  $(C_i^{(j)})_{j=1}^{L_i}$ , are sampled independent of the empirical circadian distribution  $P_C$ :

$$\mathbb{P}[C_i^{(j)} = c] = P_C(c) \quad \forall j \in \{1, \dots, L_i\}. \quad (\text{Equation 5})$$

For the spatial component, for each user, a connected sub-graph  $S_i$  of size 10 is first sampled from the Delaunay tessellation of the antenna coordinates  $\mathcal{L}$ . This sub-graph is then randomly ordered as a list, which we denote by  $S_i = (S_i^{(k)})_{k=1}^{10}$  with a slight abuse of notations. Finally, the locations of the records  $X_i^{(j)} \in X_i$  are sampled independent of  $S_i$  according to the empirical frequency distribution  $P_F$ :

$$\mathbb{P}[X_i^{(j)} = S_i^{(k)}] = P_F(k) \quad \forall j \in \{1, \dots, L_i\}. \quad (\text{Equation 6})$$

Note that when the size of the dataset  $N$  sampled by our model  $\mathcal{M}$  increases, this corresponds to sampling more individuals from the same underlying geography. This is what we mean throughout this work when we increase the size of the dataset, e.g., in unicity curves (Figure 1): we consider the dataset to be a growing sample from the same underlying population.

#### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.patter.2021.100204>.

#### ACKNOWLEDGMENTS

The authors would like to thank Shubham Jain for their comments on the codebase, and Ana-Maria Cretu, Andrea Gadotti, Shubham Jain, Thibaut Lienart, Axel Oehmichen, Luc Rocher, and Arnaud Tournier for their invaluable comments on the manuscript. We acknowledge support from the Agence Française de Développement as part of its financial assistance to the OPAL project.

#### AUTHOR CONTRIBUTIONS

A.F. designed and performed the experiments, built the models, helped with the mathematical results, and drafted the manuscript. F.H. derived the mathematical results, advised on model construction, and revised the manuscript. Y-A.d.M. designed the experiments and revised the manuscript.

#### DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: August 24, 2020  
Revised: November 27, 2020  
Accepted: January 7, 2021  
Published: February 12, 2021

#### REFERENCES

- Vodafone. (2018). Vodafone UK's company history and achievements. <https://www.vodafone.co.uk/about-us/company-history/>.
- Lomas, N. (2017). How "anonymous" wifi data can still be a privacy risk (TechCrunch). <http://tcrn.ch/2ywxGdy>.
- Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., and Tatem, A.J. (2014). Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci. U S A* 111, 15888–15893.
- Ratti, C., Frenchman, D., Pulselli, R.M., and Williams, S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. *Environ. Plann. B Plann. Des.* 33, 727–748.

- Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., and Buckee, C.O. (2012). Quantifying the impact of human mobility on malaria. *Science* 338, 267–270.
- Gomes, M.F., Pastore, Y., Piontti, A., Rossi, L., Chao, D., Longini, I., Halloran, M.E., and Vespignani, A. (2014). Assessing the International spreading risk associated with the 2014 west African Ebola outbreak. *PLoS Currents* 6, ecurrents.outbreaks.cd818f63d40e24ef769d-da7df9e0da5.
- Mari, L., Bertuzzo, E., Righetto, L., Casagrandi, R., Gatto, M., Rodriguez-turbe, I., and Rinaldo, A. (2012). Modelling cholera epidemics: the role of waterways, human mobility and sanitation. *J. R. Soc. Interface* 9, 376–388.
- Bajardi, P., Poletto, C., Ramasco, J.J., Tizzoni, M., Colizza, V., and Vespignani, A. (2011). Human mobility networks, travel restrictions, and the global spread of 2009 H1n1 pandemic. *PLoS One* 6, e16591.
- Merler, S., and Ajelli, M. (2009). The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proc. Biol. Sci.* 277, 557–565.
- Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., Gabrilovich, E., Gadepalli, K., Gipson, B., Guevara, M., et al. (2020). Google COVID-19 community mobility reports: anonymization process description (version 1.0). arXiv, preprint arXiv:2004.04145.
- Steele, J.E., Sundsøy, P.R., Pezzulo, C., Alegana, V.A., Bird, T.J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.A., Iqbal, A.M., et al. (2017). Mapping poverty using mobile phone and satellite data. *J. R. Soc. Interface* 14, 20160690.
- Toole, J.L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M.C., and Lazer, D. (2015). Tracking employment shocks using mobile phone data. *J. R. Soc. Interface* 12, 20150185.
- Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., and Buckee, C.O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface* 10, 20120986.
- Madden, M., Rainie, L., Zickuhr, K., Duggan, M., and Smith, A. (2014). Public Perceptions of Privacy and Security in the Post-Snowden Era, 12 (Pew Research Center).
- Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science* 350, 1073–1076.
- Li G., Yu L., Ng W.S., Wu W., and Goh S.T. Predicting Home and Work Locations Using Public Transport Smart Card Data by Spectral Analysis. In 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pages 2788–2793, Gran Canaria, Spain, September 2015. IEEE.
- Ashbrook D. and Starner T. Learning significant locations and predicting user movement with GPS. In Proceedings. Sixth International Symposium on Wearable Computers, pages 101–108, Seattle, WA, USA, 2002. IEEE.
- Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. In *Pervasive Computing, Volume 6696 of Lecture Notes in Computer Science*, K. Lyons, J. Hightower, and E.M. Huang, eds. (Springer Berlin Heidelberg), pp. 133–151.
- Mahmud, J., Nichols, J., and Drews, C. (2014). Home location identification of Twitter users. *ACM Trans. Intell. Syst. Technol.* 5, 47.
- Li R., Wang S., Deng H., Wang R, and Chen-Chuan Chang K. Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1023–1031, New York, NY, USA, 2012. ACM.
- Cho E., Myers S.A., and Leskovec J. Friendship and Mobility: User Movement in Location-based Social Networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1082–1090, New York, NY, USA, 2011. ACM.
- Monsivais, D., Ghosh, A., Bhattacharya, K., Dunbar, R.I.M., and Kaski, K. (2017). Tracking urban human activity from mobile phone calling patterns. *PLoS Comput. Biol.* 13, e1005824.

23. Monsivais, D., Bhattacharya, K., Ghosh, A., Dunbar, R.I.M., and Kaski, K. (2017). Seasonal and geographical impact on human resting periods. *Sci. Rep.* 7, 10717.
24. Kondor, D., Hashemian, B., de Montjoye, Y.-A., and Ratti, C. (2020). Towards matching user mobility traces in large-scale datasets. In *IEEE Transactions on Big Data*, 6, p. 1, 714-726.
25. Hasan S., Zhan X., and Ukkusuri S.V. Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing, UrbComp '13*, pages 6:1–6:8, New York, NY, USA, 2013. ACM.
26. Ahas, R., Aasa, A., Silm, S., and Tiru, M. (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. *Transport. Res. C Emerg. Tech.* 18, 45–54.
27. Felbo, B., Sundsøy, P., Pentland, A., Lehmann, S., and de Montjoye, Y.-A. (2017). Modeling the temporal nature of human behavior for demographics prediction. In *Machine learning and knowledge discovery in databases*, volume 10536 of *lecture notes in computer science* (Springer), pp. 140–152.
28. de Montjoye, Y.-A., Quoidbach, J., Robic, F., and Pentland, A.S. (2013). Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction* (Springer), pp. 48–55.
29. Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U S A* 104, 7332–7336.
30. Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., De Menezes, M.A., Kaski, K., Barabási, A.-L., and Kertész, J. (2007). Analysis of a large-scale weighted network of one-to-one human communication. *New J. Phys.* 9, 179.
31. Krumme, C., Llorente, A., Cebrian, M., Pentland, A.S., and Moro, E. (2013). The predictability of consumer visitation patterns. *Sci. Rep.* 3, 1645.
32. de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., and Blondel, V.D. (2013). Unique in the crowd: the privacy bounds of human mobility. *Sci. Rep.* 3, 1376.
33. Pellungrini, R., Pappalardo, L., Pratesi, F., and Monreale, A. (2017). A data mining approach to assess privacy risk in human mobility data. *ACM Trans. Intell. Syst. Technol.* 9, 31.
34. Achara, J.P., Acs, G., and Castelluccia, C. (2015). On the Unicity of Smartphone Applications (ACM Press), pp. 27–36.
35. Sekara, V., Mones, E., and Jonsson, H. (2018). Temporal limits of privacy in human behavior. *arXiv*, preprint arXiv:1806.03615.
36. de Montjoye, Y.-A., Radaelli, L., Singh, V.K., and Pentland, A.P. (2015). Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science* 347, 536–539.
37. Xu, Y., Belyi, A., Bojic, I., and Ratti, C. (2018). Human mobility and socio-economic status: analysis of Singapore and Boston. *Comput. Environ. Urban Syst.* 72, 51–67.
38. Deußner C., Passmann S., and Strufe T. Browsing unicity: On the limits of anonymizing web tracking data. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 777–790. IEEE, 2020.
39. Narayanan, A., and Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets (IEEE), pp. 111–125.
40. Riederer, C., Kim, Y., Chaintreau, A., Korula, N., and Lattanzi, S. (2016). Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th International Conference on World Wide Web (International World Wide Web Conferences Steering Committee)*, pp. 707–719.
41. Snchez, D., Martnez, S., and Domingo-Ferrer, J. (2016). Comment on "Unique in the shopping mall: on the reidentifiability of credit card metadata". *Science* 351, 1274.
42. El Emam, K. (2015). On Re-identification: Not Really Unique in the Shopping Mall.
43. Barth-Jones, D., El Emam, K., Bambauer, J., Cavoukian, A., and Malin, B. (2015). Assessing data intrusion threats. *Science* 348, 194–195.
44. Pappalardo, L., and Simini, F. (2018). Data-driven generation of spatio-temporal routines in human mobility. *Data Min. Knowl. Discov.* 32, 787–829.
45. Gonzalez, M.C., Hidalgo, C.A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature* 453, 779.
46. Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., and Baronchelli, A. (2018). Evidence for a conserved quantity in human mobility. *Nat. Hum. Behav.* 2, 1.
47. Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nat. Phys.* 6, 818–823.
48. Hasan, S., Schneider, C.M., Ukkusuri, S.V., and González, M.C. (April 2013). Spatiotemporal patterns of urban human mobility. *J. Stat. Phys.* 151, 304–318.
49. Article 29 Data Protection Working Party (2014). Opinion 05/2014 on Anonymisation Techniques (European Commission).
50. Centre for Humanitarian Data of the United Nations Office for the Coordination of Humanitarian Affairs (2019). Guidance Note Series on Data Responsibility on Humanitarian Action. Note 1: Statistical Disclosure Control (United Nations).
51. Gramaglia, M., and Fiore, M. (2014). On the anonymizability of mobile traffic datasets. *arXiv*, preprint arXiv:1501.00100.
52. Oehmichen, A., Jain, S., Gadotti, A., and de Montjoye, Y.-A. (2019). Opal: high performance platform for large-scale privacy-preserving location data analytics. In *2019 IEEE International Conference on Big Data (Big Data) (IEEE)*, pp. 1332–1342.
53. Mir, D.J., Isaacman, S., Cáceres, R., Martonosi, M., and Wright, R.N. (2013). Dp-where: Differentially private modeling of human mobility. In *2013 IEEE international conference on big data (IEEE)*, pp. 580–588.
54. Francis, P., Probst Eide, S., and Munz, R. (2017). Diffix: high-utility database anonymization. In *Annual Privacy Forum* (Springer), pp. 141–158.
55. Gonzalez, M.C., Hidalgo, C.A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature* 453, 779.