OPEN

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.

D erived from the Latin Privatus, meaning "withdraw from public life," the notion of privacy has been foundational to the development of our diverse societies, forming the basis for individuals' rights such as free speech and religious freedom[1]. Despite its importance, privacy has mainly relied on informal protection mechanisms. For instance, tracking individuals' movements has been historically difficult, making them de-facto private. For centuries, information technologies have challenged these informal protection mechanisms. In 1086, William I of England commissioned the creation of the Doomsday book, a written record of major property holdings in England containing individual information collected for tax and draft purposes[2]. In the late 19th century, de-facto privacy was similarly threatened by photographs and yellow journalism. This resulted in one of the first publications advocating privacy in the U.S. in which Samuel Warren and Louis Brandeis argued that privacy law must evolve in response to technological changes[3].

Modern information technologies such as the Internet and mobile phones, however, magnify the uniqueness of individuals, further enhancing the traditional challenges to privacy. Mobility data is among the most sensitive data currently being collected. Mobility data contains the approximate whereabouts of individuals and can be used to reconstruct individuals' movements across space and time. Individual mobility traces T [Fig. 1A–B] have been used in the past for research purposes[4–18] and to provide personalized services to users[19]. A list of potentially sensitive professional and personal information that could be inferred about an individual knowing only his mobility trace was published recently by the Electronic Frontier Foundation[20]. These include the movements of a competitor sales force, attendance of a particular church or an individual's presence in a motel or at an abortion clinic.

While in the past, mobility traces were only available to mobile phone carriers, the advent of smartphones and other means of data collection has made these broadly available. For example, Apple® recently updated its privacy policy to allow sharing the spatio-temporal location of their users with "partners and licensees"[21]. 65.5B geo-tagged payments are made per year in the US[22] while Skyhook wireless is resolving 400 M user's WiFi location every day[23]. Furthermore, it is estimated that a third of the 25B copies of applications available on Apple's App Store[SM] access a user's geographic location[24,25], and that the geo-location of ~50% of all iOS and Android traffic is available to ad networks[26]. All these are fuelling the ubiquity of simply anonymized mobility datasets and are giving room to privacy concerns.

A simply anonymized dataset does not contain name, home address, phone number or other obvious identifier. Yet, if individual's patterns are unique enough, outside information can be used to link the data back to an individual. For instance, in one study, a medical database was successfully combined with a voters list to extract
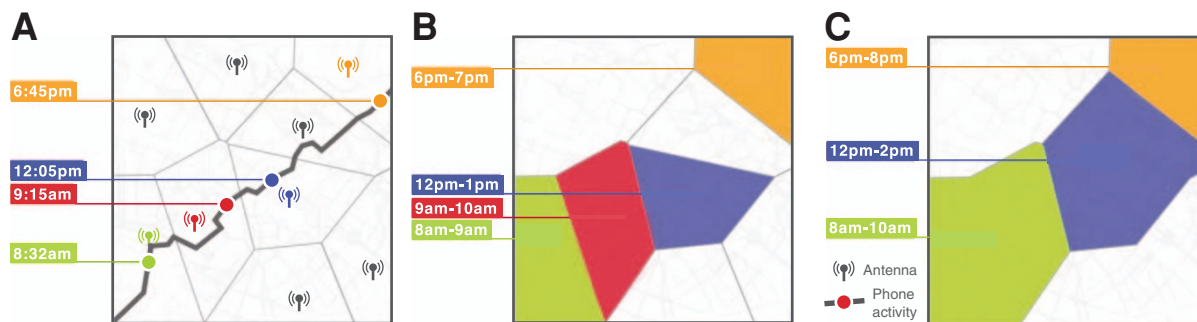
Figure 1 | (A) Trace of an anonymized mobile phone user during a day. The dots represent the times and locations where the user made or received a call. Every time the user has such an interaction, the closest antenna that routes the call is recorded. (B) The same user's trace as recorded in a mobility database. The Voronoi lattice, represented by the grey lines, are an approximation of the antennas reception areas, the most precise location information available to us. The user's interaction times are here recorded with a precision of one hour. (C) The same individual's trace when we lower the resolution of our dataset through spatial and temporal aggregation. Antennas are aggregated in clusters of size two and their associated regions are merged. The user's interaction are recorded with a precision of two hours. Such spatial and temporal aggregation render the 8:32 am and 9:15 am interactions indistinguishable.

the health record of the governor of Massachusetts[27]. In another, mobile phone data have been re-identified using users' top locations[28]. Finally, part of the Netflix challenge dataset was re-identified using outside information from The Internet Movie Database[29].

All together, the ubiquity of mobility datasets, the uniqueness of human traces, and the information that can be inferred from them highlight the importance of understanding the privacy bounds of human mobility. We show that the uniqueness of human mobility traces is high and that mobility datasets are likely to be re-identifiable using information only on a few outside locations. Finally, we show that one formula determines the uniqueness of mobility traces providing mathematical bounds to the privacy of mobility data. The uniqueness of traces is found to decrease according to a power function with an exponent that scales linearly with the number of known spatio-temporal points. This implies that even coarse datasets provide little anonymity.

Results

Uniqueness of human mobility. In 1930, Edmond Locard showed that 12 points are needed to uniquely identify a fingerprint[30]. Our unicity test estimates the number of points p needed to uniquely identify the mobility trace of an individual. The fewer points needed, the more unique the traces are and the easier they would be to re-identify using outside information. For re-identification purposes, outside observations could come from any publicly available information, such as an individual's home address, workplace address, or geo-localized tweets or pictures. To the best of our knowledge, this is the first quantification of the uniqueness of human mobility traces with random points in a sparse, simply anonymized mobility dataset of the scale of a small country.

Given I_p, a set of spatio-temporal points, and D, a simply anonymized mobility dataset, we evaluate ε, the uniqueness of traces, by extracting from D the subset of trajectories $S(I_p)$ that match the p points composing I_p [See Methods]. A trace is unique if $|S(I_p)| = 1$, containing only one trace. For example, in Fig. 2A, we evaluate the uniqueness of traces given $I_{p=2}$. The two spatio-temporal points contained in $I_{p=2}$ are zone I from 9am to 10am and zone II from 12pm to 1pm. The red and the green traces both satisfy $I_{p=2}$, making them not unique. However, we can also evaluate the uniqueness of traces knowing $I_{p=3}$, adding as a third point zone III between 3pm and 4pm. In this case $|S(I_{p=3})| = 1$, uniquely characterize the green trace. A lower bound on the risk of deductive disclosure of a user's identity is given by the uniqueness of his mobility trace, the likelihood of this brute force characterization to succeed.

Our dataset contains 15 months of mobility data for 1.5 M people, a significant and representative part of the population of a small European country, and roughly the same number of users as the

location-based service Foursquare®[31]. Just as with smartphone applications or electronic payments, the mobile phone operator records the interactions of the user with his phone. This creates a comparable longitudinally sparse and discrete database [Fig. 3]. On average, 114 interactions per user per month for the nearly 6500 antennas are recorded. Antennas in our database are distributed throughout the country and serve, on average, ~ 2000 inhabitants each, covering areas ranging from 0.15 km² in cities to 15 km² in rural areas. The number of antennas is strongly correlated with population density ($R^2 = .6426$) [Fig. 3C]. The same is expected from businesses, places in location-based social networks, or WiFi hotspots.

Fig. 2B shows the fraction of unique traces (ε) as a function of the number of available points p. Four randomly chosen points are enough to uniquely characterize 95% of the users ($\varepsilon > .95$), whereas two randomly chosen points still uniquely characterize more than 50% of the users ($\varepsilon > .5$). This shows that mobility traces are highly unique, and can therefore be re-identified using little outside information.

Scaling properties. Nonetheless, ε depends on the spatial and temporal resolution of the dataset. Here, we determine this dependence by lowering the resolution of our dataset through spatial and temporal aggregation [Fig. 1C]. We do this by increasing the size of a region, aggregating neighbouring cells into clusters of v cells, or by reducing the dataset's temporal resolution, increasing the length of the observation time window to h hours [see Methods]. Both of these aggregations are bound to decrease ε, and therefore, make re-identification harder.

Fig. 4A shows how the uniqueness of mobility traces ε depends on the spatial and temporal resolution of the data. This reduction, however, is quite gradual. Given four points (p=4), we find that $\varepsilon > .5$ when using a resolution of $h = 5$ hours and $v = 5$ antennas.

Statistically, we find that traces are more unique when coarse on one dimension and fine along another than when they are medium-grained along both dimensions. Indeed, given four points, $\varepsilon > .6$ in a dataset with a temporal resolution of $h = 15$ hours or a spatial resolution of $v = 15$ antennas while $\varepsilon > .4$ in a dataset with a temporal resolution of $h = 7$ hours and a spatial resolution of $v = 7$ antennas [Fig. 4A].

Next, we show that it is possible to find one formula to estimate the uniqueness of traces given both, the spatial and temporal resolution of the data, and the number of points available to an outside observer. Fig. 4B and 4C show that the uniqueness of a trace decreases as the power function $\varepsilon = \alpha - x^\beta$, for decreases in both the spatial and temporal resolution (x), and for all considered $p = 4, 6, 8$ and 10 (see Table S1). The uniqueness of human mobility can thus be expressed using the single formula: $\varepsilon = \alpha - (vh)^\beta$. We find that this power
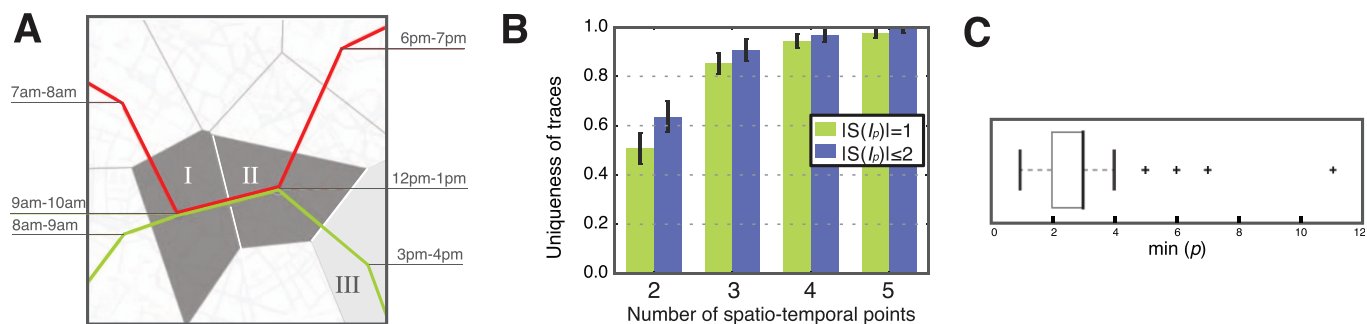
Figure 2 | (A) $I_{p=2}$ means that the information available to the attacker consist of two 7am-8am spatio-temporal points (I and II). In this case, the target was in zone I between 9am to 10am and in zone II between 12pm to 1pm. In this example, the traces of two anonymized users (red and green) are compatible with the constraints defined by $I_{p=2}$. The subset $S(I_{p=2})$ contains more than one trace and is therefore not unique. However, the green trace would be uniquely characterized if a third point, zone III between 3pm and 4pm, is added ($I_{p=3}$). (B) The uniqueness of traces with respect to the number p of given spatio-temporal points (I_p). The green bars represent the fraction of unique traces, i.e. $|S(I_p)| = 1$. The blue bars represent the fraction of $|S(I_p)| \leq 2$. Therefore knowing as few as four spatio-temporal points taken at random ($I_{p=4}$) is enough to uniquely characterize 95% of the traces amongst 1.5 M users. (C) Box-plot of the minimum number of spatio-temporal points needed to uniquely characterize every trace on the non-aggregated database. At most eleven points are enough to uniquely characterize all considered traces.

function fits the data better than other two-parameters functions such as $\alpha - \exp(\lambda x)$, a stretched exponential $\alpha - \exp x^{\beta}$, or a standard linear function $\alpha - \beta x$ (see Table S1). Both estimators for α and β are highly significant ($p < 0.001$)[32], and the mean pseudo-R^2 is 0.98 for the $I_{p=4}$ case and the $I_{p=10}$ case. The fit is good at all levels of spatial and temporal aggregation [Fig. S3A–B].

The power-law dependency of ε means that, on average, each time the spatial or temporal resolution of the traces is divided by two, their uniqueness decreases by a constant factor $\sim (2)^{-\beta}$. This implies that privacy is increasingly hard to gain by lowering the resolution of a dataset.

Fig. 2B shows that, as expected, ε increases with p. The mitigating effect of p on ε is mediated by the exponent β which decays linearly with p: $\beta = 0.157 - 0.007p$ [Fig. 4E]. The dependence of β on p implies that a few additional points might be all that is needed to identify an individual in a dataset with a lower resolution. In fact, given four points, a two-fold decrease in spatial or temporal resolution makes it 9.3% less likely to identify an individual, while given ten points, the same two-fold decrease results in a reduction of only 6.2% (see Table S1).

Because of the functional dependency of ε on p through the exponent β, mobility datasets are likely to be re-identifiable using information on only a few outside locations.

Discussion

Our ability to generalize these results to other mobility datasets depends on the sensitivity of our analysis to extensions of the data to larger populations, or geographies. An increase in population density will tend to decrease ε. Yet, it will also be accompanied by an increase in the number of antennas, businesses or WiFi hotspots used for localizations. These effects run opposite to each other, and therefore, suggest that our results should generalize to higher population densities.

Extensions of the geographical range of observation are also unlikely to affect the results as human mobility is known to be highly circumscribed. In fact, 94% of the individuals move within an average radius of less than 100 km[17]. This implies that geographical extensions of the dataset will stay locally equivalent to our observations, making the results robust to changes in geographical range.

From an inference perspective, it is worth noticing that the spatio-temporal points do not equally increase the likelihood of uniquely identifying a trace. Furthermore, the information added by a point is highly dependent from the points already known. The amount of information gained by knowing one more point can be defined as the reduction of the cardinality of $S(I_p)$ associated with this extra point. The larger the decrease, the more useful the piece of information is. Intuitively, a point on the MIT campus at 3AM is more likely to make a trace unique than a point in downtown Boston on a Friday evening.

This study is likely to underestimate ε, and therefore the ease of re-identification, as the spatio-temporal points are drawn at random from users' mobility traces. Our I_p are thus subject to the user's spatial and temporal distributions. Spatially, it has been shown that the uncertainty of a typical user's whereabouts measured by its
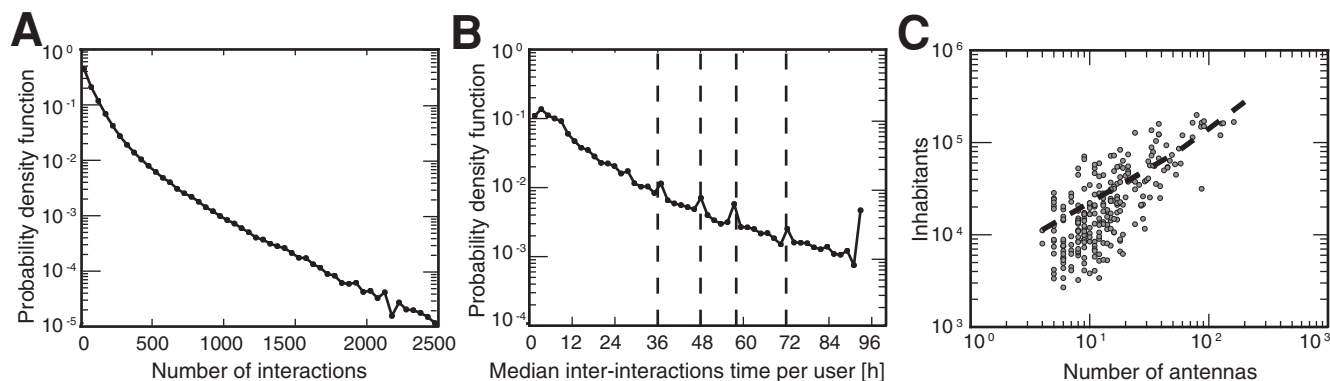
Figure 3 | (A) Probability density function of the amount of recorded spatio-temporal points per user during a month. (B) Probability density function of the median inter-interaction time with the service. (C) The number of antennas per region is correlated with its population ($R^2 = .6426$). These plots strongly emphasize the discrete character of our dataset and its similarities with datasets such as the one collected by smartphone apps.
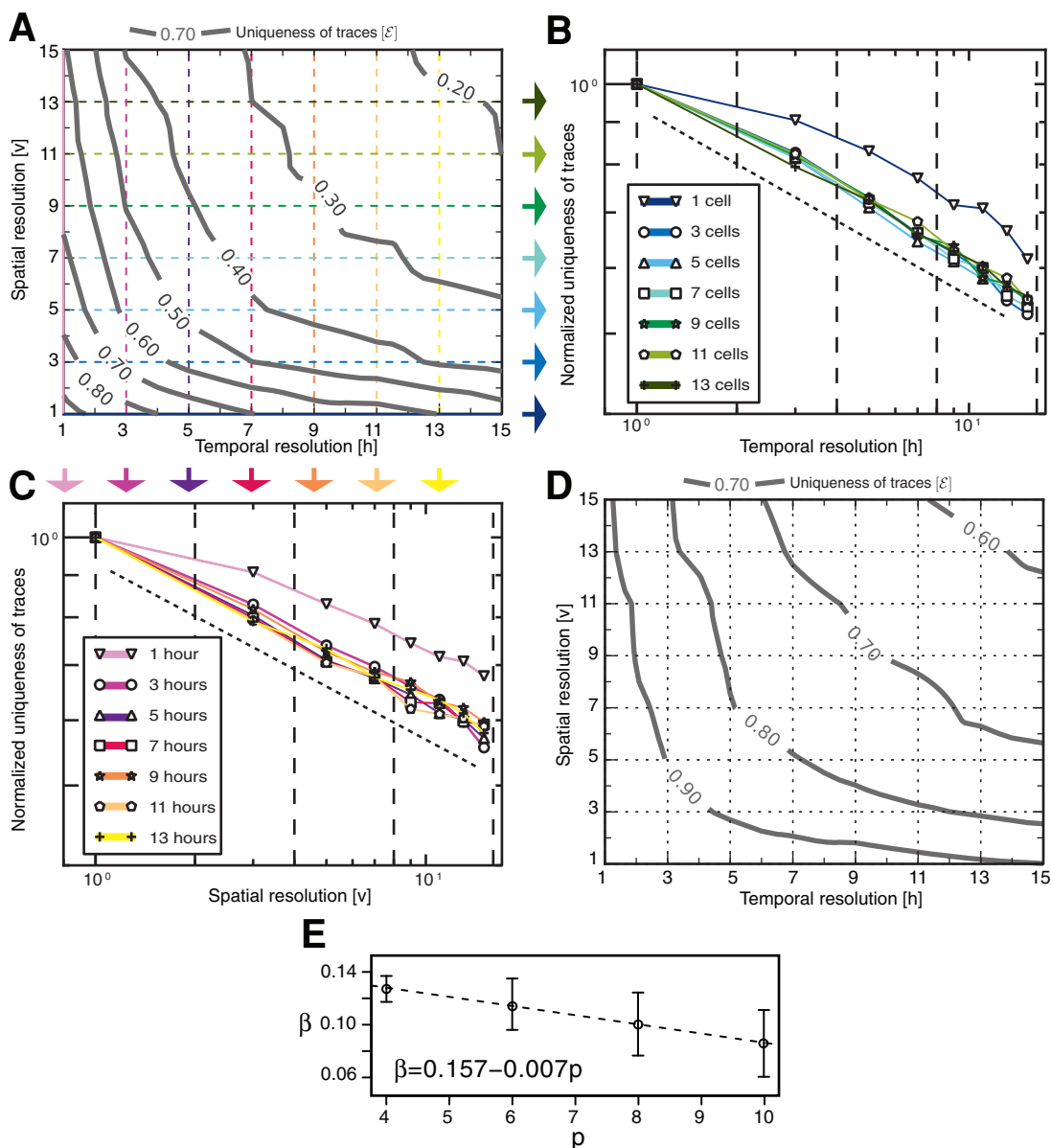
Figure 4 | Uniqueness of traces [ε] when we lower the resolution of the dataset with (A) $p = 4$ and (D) $p = 10$ points. It is easier to attack a dataset that is coarse on one dimension and fine along another than a medium-grained dataset along both dimensions. Given four spatio-temporal points, more than 60% of the traces are uniquely characterized in a dataset with an $h = 15$-hours temporal resolution while less than 40% of the traces are uniquely characterized in a dataset with a temporal resolution of $h = 7$ hours and with clusters of $v = 7$ antennas. The region covered by an antenna ranges from 0.15 km² in urban areas to 15 km² in rural areas. (B–C) When lowering the temporal or the spatial resolution of the dataset, the uniqueness of traces decrease as a power function $\varepsilon = \alpha - x^{\beta}$. (E) While ε decreases according to a power function, its exponent β decreases linearly with the number of points p. Accordingly, a few additional points might be all that is needed to identify an individual in a dataset with a lower resolution.

entropy is 1.74, less than two locations[18]. This makes our random choices of points likely to pick the user's top locations (typically "home" and "office"). Temporally, the distribution of calls during the week is far from uniform [Fig. S1] which makes our random choice more likely to pick a point at 4PM than at 3AM. However, even in this case, the traces we considered that are most difficult to identify can be uniquely identified knowing only 11 locations [Fig. 2C].

For the purpose of re-identification, more sophisticated approaches could collect points that are more likely to reduce the uncertainty, exploit irregularities in an individual's behaviour, or implicitly take into account information such as home and work-place or travels abroad[29,33]. Such approaches are likely to reduce the number of locations required to identify an individual, vis-à-vis the average uniqueness of traces.

We showed that the uniqueness of human mobility traces is high, thereby emphasizing the importance of the idiosyncrasy of human movements for individual privacy. Indeed, this uniqueness means that little outside information is needed to re-identify the trace of a targeted individual even in a sparse, large-scale, and coarse mobility dataset. Given the amount of information that can be inferred from mobility data, as well as the potentially large number of simply anonymized mobility datasets available, this is a growing concern. We further showed that while $\mathcal{E} \sim (vh)^{\beta}$, $\beta \sim -p/100$. Together, these determine the uniqueness of human mobility traces given the traces' resolution and the available outside information. These results should inform future thinking in the collection, use, and protection of mobility data. Going forward, the importance of location data will only increase[34] and knowing the bounds of individual's privacy will

be crucial in the design of both future policies and information technologies.

Methods

The dataset. This work was performed using an anonymized mobile phone dataset that contains call information for ~1.5 M users of a mobile phone operator. The data collection took place from April 2006 to June 2007 in a western country. Each time a user interacts with the mobile phone operator network by initiating or receiving a call or a text message, the location of the connecting antenna is recorded [Fig. 1A]. The dataset's intrinsic spatial resolution is thus the maximal half-distance between antennas. The dataset's intrinsic temporal resolution is one hour [Fig. 1B].

Unicity test and the likelihood of deductive disclosure. The considered dataset contains one trace T for each user. The traces spatio-temporal points contain the region in which the user was and the time of the interaction. We evaluate the uniqueness of a trace given a set I_p of p randomly chosen spatio-temporal points. A trace is said to be to be compatible with I_p if $I_p \subseteq T$ [Fig. 2A]. Note that this notion of compatibility can easily be extended to noisier or richer data. A brute force characterization is performed by extracting from the entire dataset of 1.5 M users $S(I_p)$, the set of users whose mobility traces T are compatible with I_p. All mobility traces in the dataset T are successively tested for compatibility with I_p. A trace is characterized "out of x", if the set of traces that are compatible with the points contains at most x users: $|S(I_p)| \leq x$. A trace is uniquely characterized if the set contains exactly one trace: $|S(I_p)| = 1$. The uniqueness of traces is estimated as the percentage of 2500 random traces that are unique given p spatio-temporal points. The p points composing I_p are taken at random among all the interactions the user had with the service. As discussed, we do not apply any constraints regarding the choice of I_p.

Minimum number of spatio-temporal location needed to uniquely characterize every trace. Fig. 2B shows that $.95 < \varepsilon < 1$ given $I_{p=4}$. Fig. 2C evaluates the minimum p needed to uniquely characterize every trace in a given set. This set contains a random sample of 1000 heavy-users, i.e. users that used their phone at least 75 times per month as their randomly chosen points might make their trace less unique.

Spatial aggregation. Spatial aggregation is achieved by increasing the size of the regions in which the user is known to be during his interactions with the service. In the case of discrete data, a bijective relation exists between antennas (known in this case as centroids) and the region defined by the Voronoi tessellation. The tessellation is defined so that every point in a region is closer to the region's antenna than to any other antenna. In order to increase the region's area, one should group antennas into clusters of a given size v. While the problem of optimally grouping places in a 2D space into groups of given sizes v is non trivial, it can be approximated through clustering methods. The canonical clustering methods focus on minimizing the within-cluster sum of squares rather than producing balanced clusters. This drawback can be controlled by the use of a Frequency Sensitive Competitive Learning scheme[35]. Fig. S2 shows the resulting group size histogram optimized for clusters of size 4. Once antennas are aggregated into groups, their associated regions are merged.

1. Clippinger, J. In *Rules for Growth: Promoting Innovation and Growth Through Legal Reform* (Kauffman Foundation, Kansas City, 2010).
2. Clanchy, M. T. *From Memory to Written Records England 1066–1307* (Harvard University Press, Cambridge, 1979).
3. Warren, S. & Brandeis, L. The right to privacy. *Harvard Law Review* **193**, 193–220 (1890).
4. Hey, T., Tansley, S. & Tolle, K. (eds) *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, Redmond, 2009).
5. Barabasi, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).
6. Clauset, A. & Eagle, N. Persistence and periodicity in a dynamic proximity network. *Proc. DIMACS* (2007).
7. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**, 1029–1031 (2010).
8. Eagle, N., Pentland, A. & Lazer, D. Inferring social network structure using mobile phone data. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 15274–15278 (2009).
9. Eagle, N., de Montjoye, Y.-A. & Bettencourt, L. Community computing: Comparisons between rural and urban societies using mobile phone data. *Computational Science and Engineering* **4**, 144–150 (2009).
10. Reader, T. *et al.* Predictors of short-term decay of cell phone contacts in a large scale communication network. *Social Networks* **33**, 245–257 (2011).
11. Hidalgo, C. & Rodriguez, C. The dynamics of a mobile phone network. *Physica A* **387**, 3017–3024 (2008).
12. Newman, M. E. J. *Networks: An Introduction* (Oxford University Press, New York, 2010).
13. Onnela, J.-P. *et al.* Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7332–7336 (2007).
14. Szell, M., Lambiotte, R. & Thurner, S. Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 13636–13641 (2010).
15. Meloni, S. *et al.* Modeling human mobility responses to the large-scale spreading of infectious diseases. *Sci. Rep.* **1**, 62 (2011).
16. Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21484–21489 (2009).
17. Gonzalez, M., Hidalgo, C. & Barabasi, A. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
18. Song, C., Qu, Z., Blumm, N. & Barabasi, A. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).
19. Finding places on the go has never been easier. http://blog.foursquare.com/2012/02/08/finding-places-on-the-go-has-never-been-easier-%E2%80%93-check-out-the-new-explore-for-your-phone/, Accessed 2012 Jul. 1.
20. Blumberg, A. & Eckersley, P. On locational privacy and how to avoid losing it forever. *E.F.F.* (2009).
21. Apple privacy policy, http://www.apple.com/legal/privacy/, Accessed 2011 Jul. 25.
22. Federal reserve financial services federal reserve study shows more than three-quarters of non-cash payments are now electronic. *Federal Reserve* (2010).
23. Skyhook wireless spotRank overview, Available: http://www.skyhookwireless.com/location-intelligence/, Accessed 2012 Jul. 17.
24. Apples app store downloads top 25 billion, http://www.apple.com/pr/library/2012/03/05Apples-App-Store-Downloads-Top-25-Billion.html, Accessed 2012 Mar. 28.
25. The app genome project. http://blog.myLookout.com/, Accessed 2011 Jul. 27.
26. Mobile geo-location advertising will be a big number in 2015. http://adfonic.com/wp-content/uploads/2012/03/geo-location-white-paper.pdf, Accessed 2012 Jul. 17.
27. Sweeney, L. k-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness and Knowledge-Based Systems* **10**, 557–570 (2002).
28. Zang, H. & Bolot, J. Anonymization of location data does not work: A large-scale measurement study. *Proc. Int. Conf. on Mobile computing and networking* **17**, 145–156 (2011).
29. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. *IEEE Trans. Secur. Priv.* **8**, 111–125 (2008).
30. Locard, E. *Traité de criminalistique.* (J. Desvigne et ses fils Lyon, 1931).
31. Boom! Foursquare crosses 2 million users. http://techcrunch.com/2010/07/10/foursquare-crosses-2-million-users/, Accessed 2010 Aug. 25.
32. Bates, D. & Watts, D. *Nonlinear Regression Analysis and Its Applications* (Wiley, Hoboken, 1988).
33. Golle, P. & Partridge, K. On the anonymity of home/work location pairs. *Pervasive Computing* 390–397 (2009).
34. Manyika, J. *et al.* Big data: The next frontier for innovation, competition and productivity. *McKinsey Global Institute* (2011).
35. Grossberg, S. Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biol. Cybern.* **23**, 121–134 (1976).

Acknowledgements

Author contributions

Y.-A. de M. designed and performed experiments, analyzed data and wrote the paper; C.A.H. designed experiments, developed analytic tools and wrote the paper; M.V. and V.D.B. designed experiments and wrote the paper.

Additional information