www.sciencemag.org/content/347/6221/536/suppl/DC1

Supplementary Materials for

Unique in the shopping mall: On the reidentifiability of credit card metadata

Yves-Alexandre de Montjoye,* Laura Radaelli, Vivek Kumar Singh, Alex "Sandy" Pentland

*Corresponding author. E-mail: yvesalexandre@demontjoye.com

Published 30 January 2015, *Science* **347**, 536 (2015)
DOI: 10.1126/science.1256297

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/347/6221/536/suppl/DC1)

Subsampled Data (Excel file)

Supplementary Materials:

Materials and Methods.

The dataset. This study was performed on an anonymized financial dataset of credit card transactions for ~1.1M people in an OECD country. The financial data along with individual gender (24% women) and income level (39% low, 35% medium, 22% high, 4% unknown) was provided to us by a major bank active in the region. The threshold between low and medium income is approximately the median household income in the country while the threshold between medium and high income is approximately 2.5 times the median household income. The data collection took place from January 1 to March 31. The median (resp. first and third quartile) of the number of transactions of people with at least one transaction every month is 8 (resp. 5 and 14). We report prices into dollars equivalent and we eliminate from the dataset 138 transactions whose price is higher than $22,800. These would make a user unique with very few points and removing them only decreases unicity. The unicity calculation [Algorithm S2] requires the entire set of raw data points for every individual. For contractual and privacy reasons, we unfortunately cannot make this raw data available. Upon request we can however make individual level data of gender, income level, resolution (h, v, a), and unicity (true, false) along with the appropriate documentation available for replication. This allows the recreation of Fig. 2, 3 and 4, as well as the GLM model and all the unicity statistics. A randomly subsampled dataset for the 4 points case can be found at http://web.media.mit.edu/~yva/uniqueintheshoppingmall/

Spatial resolution. The basic spatial resolution of the dataset is the location of the shop where the transaction took place. We decrease the spatial resolution of the data by grouping shops according to their location using a clustering algorithm. While traditional clustering aims at grouping data using a distance-based metric, Frequency-Sensitive Competitive Learning (*35*) also produces clusters of roughly the same size.

In short, in Frequency-Sensitive Competitive Learning, the chances of a cluster to win a new data point are inversely proportional–although not directly–to previous wins. This allows the algorithm to maintain a balance between clusters so that all the clusters get a similar share of the data. We here group shops using a Frequency-Sensitive Competitive Learning algorithm with μ as the number of shops that each cluster should aim to contain. Fig. S5 shows an example of the distribution of shops into clusters when the algorithm is run with parameter $\mu = 6$.

Price resolution. The dataset contains the exact price of each transaction but, as described in the manuscript, we assume that we only observe an approximation of this price with a precision *a* we call price resolution. Prices are grouped in bins whose size is increasing, i.e. the size of a bin containing low prices is smaller than the size of a bin containing high prices. For instance, a $5.33 transaction falls in the]1.8, 5.4] bin while a $35.81 transaction falls in the]16.2, 48.6] bin.

The size of bins is a function of the price resolution *a* and of the median price *m* of the bin, $m \pm (m \cdot a)$. We create bins incrementally starting from a bin centered around .4. Algorithm S1 describes in pseudo-code how we iterate from there. The algorithm has one parameter, the price resolution *a* and the algorithm terminates when the maximum price $22,800 is reached.

We report our price bins for $a=.50$ and $a=.75$ in Table S2A–B in dollars equivalent and with rounded boundaries for simplicity. We use bins computed in the original currency, and we use floating numbers in our implementation.

Unicity estimation. We estimate the value of unicity ε_p of a dataset by performing a unicity test on t=10,000 sampled users with at least p points, as described in (*19*). For each test we sampled without replacement a set of p points from the user's trace. The test is positive and the user is said to be unique if he is the only user in the entire dataset whose trace contains the p points. The unicity of the dataset is estimated as the percentage of tests that resulted in a unique trace.

$$\varepsilon_p = |\{u \in users : |S(I_p)=1| \text{ for } I_p \leftarrow \text{draw}(u,p)\}| \,/\, |users|$$

A pseudo-code for the estimation of the unicity of a dataset is given in Algorithm S2 and takes as input the number of points p. This estimation does not consider an individual gender or income level to be known; this would only increase unicity. Given a dataset D of financial traces of users, we call *trace* a sequence of points where the user was, I_p the set of points drawn from a user's trace, and $S(I_p)$ the set of traces containing I_p.

Average unicity. $\langle\Delta\varepsilon\rangle$ quantifies how much adding a dimension to the data increases unicity. We compute the average unicity $\langle\Delta\varepsilon\rangle$ at different resolutions of space and time over the linearly interpolated surface to avoid effects of sampling. It is interesting to notice in Fig. S2 that the biggest gain in unicity is achieved in the central region, where data along one dimension is high resolution and data along the other dimension is low resolution, or where data along both dimensions have a medium-grain resolution. We can also see that while adding the price of the transaction does not really help overcome a low temporal resolution (e.g. at $h=13$, $v=50$), it does help overcome a low spatial resolution (e.g. at $h=3$, $v=300$). This is likely to be because most of the transactions of a shop fall in a few bins. The transactions of a coffee shop will fall in the]2, 5] or the]5, 16] bins while the transactions of a shoe shop will fall in the]49, 146] or]146, 437] bins. Indeed, when the prices are binned at $a=.75$, the average entropy of prices per shop is $S=.31$. This is very low and means that, if we had 3 bins, 96% of the transactions would be in one bin and only 4% of the transactions would fall in the other two bins. This emphasizes the need for further computational privacy research to understand the determinants of unicity of a dataset

GLM. We use one Generalized Linear Model with a logit link function to estimate the effect of gender and income on unicity where we control for h, v, and a as factors. We used 10,000 samples per v-h-a-levels and 504 levels. All coefficients (h, v, a) are significant ($p<0.001$).

Linear Discriminant Analysis. While a full causal analysis or investigation of the determinant of re-identification of an individual are beyond the scope of this paper, we investigate potential variables through which gender or income could influence ε; the number of transactions an individual made, the number of shops or the entropy of the shops she or he went to, the number or the entropy of price bins the items she or he bought felt into ($a=.50$ and $a=.75$). We use a linear discriminant analysis with either gender or income as dependent variable and the potential variables as independent variables. For both gender and income, the entropy of the shops is the most discriminative variable.

Credit Card and Mobile Phone Records Distributions. Figure S3 shows that the behavior recorded by credit cards is very different from the one recorded by mobile phones. For example, while the use of mobile phones drops during the weekend the use of credit card strongly increases. We can also see e.g. that the use of credit cards increases steadily throughout the day until approximately 6-7pm while the use of mobile phones drops in the middle of the day during lunch hours and then peaks at approximately the same time as the use of credit cards. Finally, while the use of mobile phones peaks on Thursdays, the use of credit cards is constant across weekdays.

Figures

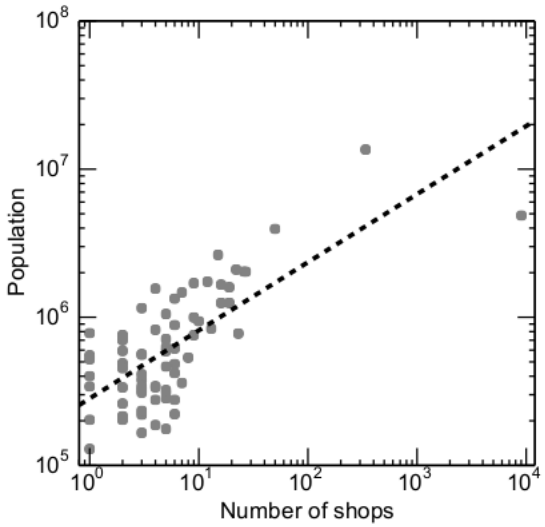

Fig. S1. The number of shops per district is strongly correlated with its population ($r^2=0.51$, $p < 0.001$). This emphasizes our ability to generalize these results to other financial datasets.

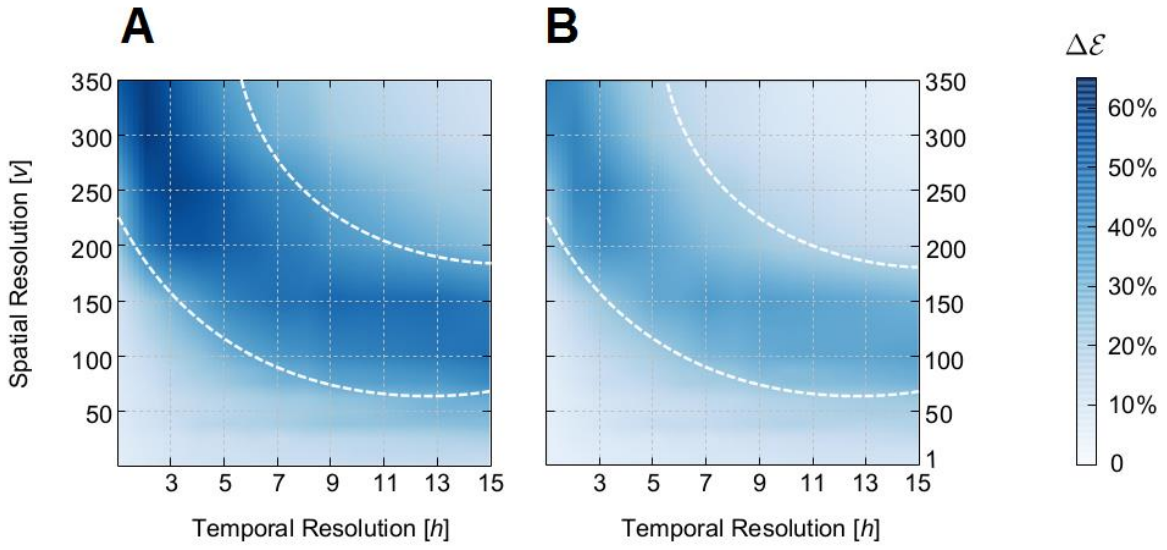

Fig. S2. Gain in unicity ($\Delta\varepsilon$) when adding a third dimension, the approximate price of a transaction (**A**, $a=.50$; **B**, $a=.75$). We see that the gain in unicity $\Delta\varepsilon$ is higher in the central region marked with dashed lines.

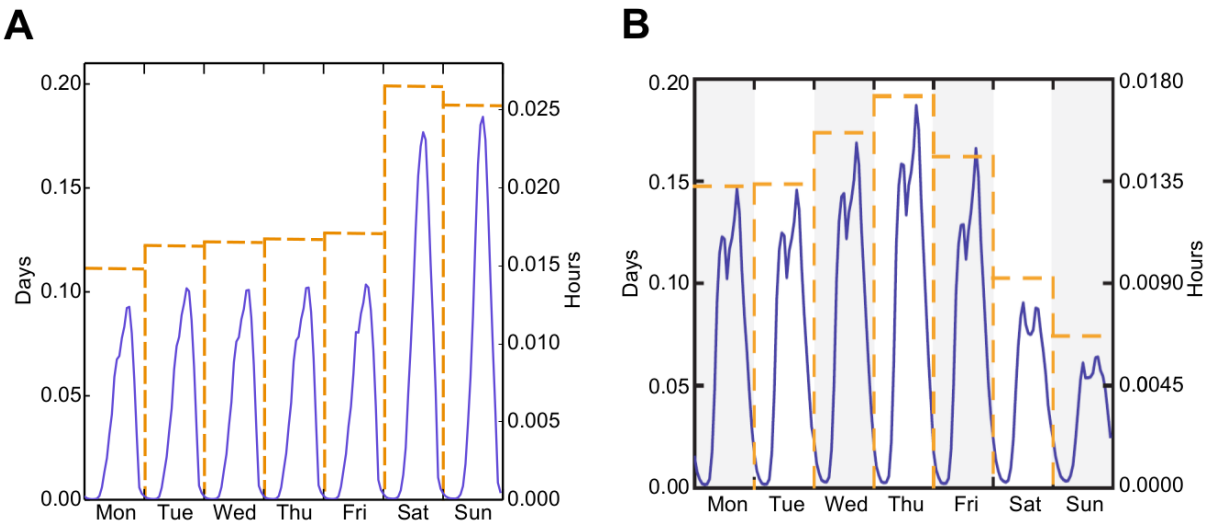

Fig. S3. (A) Probability of having a credit card record per hour (blue, right axis) and per day (orange, left axis). **(B)** Probability of having a mobile phone record per hour (blue, right axis) and per day (orange, left axis) as reported in (*19*).

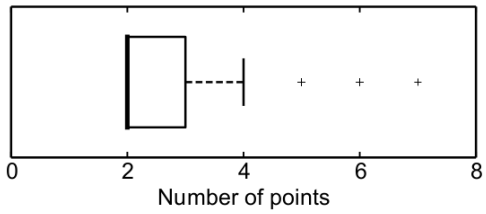

Fig. S4. While a trace may not be uniquely re-identified with *p* spatio-temporal-price triples, the same trace might be unique if more triples are known. We here evaluate the minimum number of triples *p* needed to uniquely characterize every trace in a set of 10,000 randomly sampled traces with at least *p* points (*h=1*, *v=1*, *a=.50*). In this set of traces, 7 spatio-temporal-price points are enough to re-identify all of them including the most difficult one.

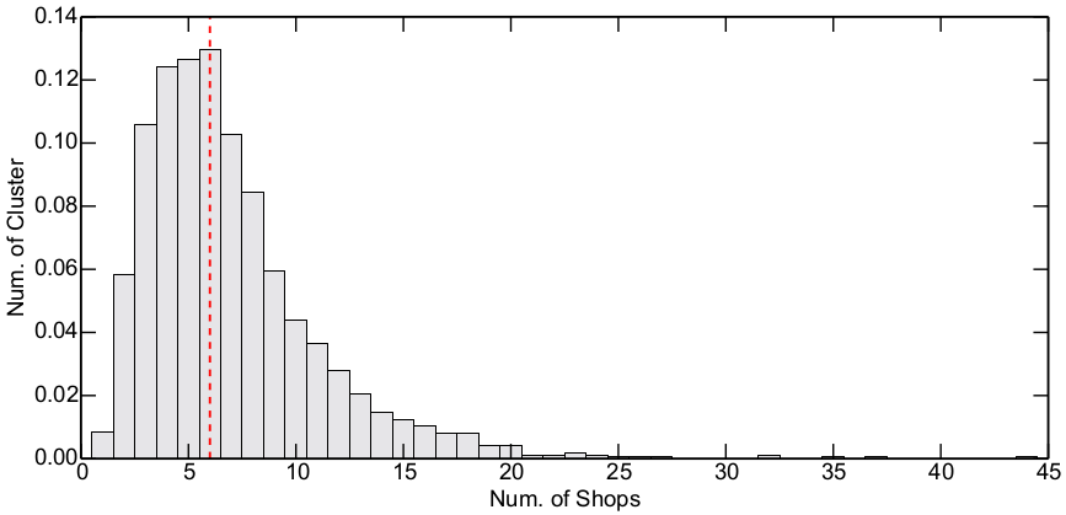

Fig. S5. Cluster's size resulting from the F.S.C.L. algorithm with $\mu = 6$. The dashed red line indicates the empirical mean of *5.9975*.

	Price Resolution [a]		
	.50	**.75**	**no price**
ε_4	.13	.06	.00
ε_6	.40	.25	.03
ε_{10}	.86	.72	.21

Table S1. Unicity at very low spatio-temporal resolution (*h=15*, *v=350*) knowing four (ε_4), six (ε_6), and ten (ε_{10}) points.

A

Bin #	Range
0	]0.2, 0.6]
1	]0.6, 1.8]
2	]1.8, 5.4]
3	]5.4, 16.2]
4	]16.2, 48.6]
5	]48.6, 145.8]
6	]145.8, 437.4]
7	]437.4, 1312.2]
8	]1312.2, 3936.6]
9	]3936.6, 11809.8]
10	]11809.8, 35429.4]

B

Bin #	Range
0	]0.1, 0.7]
1	]0.7, 4.9]
2	]4.9, 34.3]
3	]34.3, 240.1]
4	]240.1, 1680.7]
5	]1680.7, 11764.9]
6	]11764.9, 82354.3]

Table S2. (A) Bins for *a=.50*. **(B)** Bins for *a=.75*.

Algorithm S1 Bins(a)

```
top ← .4 + (.4 · a)
bins ← {.4 – (.4 · a), top}
while top ≤ 22800 do
    bottom ← top
    m ← bottom / (1 – a)
    top ← m · (1 + a)
    bins ← bins + {top}
return bins
```

Algorithm S1.

Algorithm S2 Unicity Estimation(p)

```
users ← select(D, p, 10000)
for u ∈ users do
    I_p ← draw(u, p)
    is_unique ← true
    for x ∈ D \ {u} do
        if I_p ⊂ x.trace then
            is_unique ← false
            break
    if is_unique then
        uniqueUsers ← uniqueUsers + {u}
return |uniqueUsers| / |users|
```

Algorithm S2.

Subsampled Data

This data has been made available for **replication purposes only** and cannot be redistributed.

The *xlsx* file contains the anonymized and randomly subsampled result of unicity test in the 4 points cases at each (*h*,*v*,*a*) resolution with:

- *h*: the temporal resolution
- *v*: the spatial resolution
- *a*: the price resolution
- *unique*: Whether the individual has been re-identified
- *gender*: the gender of the individual (0 for female, 1 for male)
- *income*: the income level of the individual (A for low, B for medium, C for high)

The first sheet of the spreadsheet contains the data from (1,1,0) to (15,75,75) and the second sheet contains the data from (1,87,0) to (15,350,75).

References and Notes

1. S. Higginbotham, “For science, big data is the microscope of the 21st century” (2011); http://gigaom.com/2011/11/08/for-science-big-data-is-the-microscope-of-the-21st-century/.

2. D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, Computational social science. *Science* **323**, 721–723 (2009). Medline doi:10.1126/science.1167742

3. J. Giles, Computational social science: Making the links. *Nature* **488**, 448–450 (2012). Medline doi:10.1038/488448a

4. D. J. Watts, Computational social science: Exciting progress and future directions. *Winter Issue of The Bridge on Frontiers of Engineering* **43**, 5–10 (2013).

5. A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, C. O. Buckee, Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270 (2012). Medline doi:10.1126/science.1223467

6. S. Charaudeau, K. Pakdaman, P.-Y. Boëlle, Commuter mobility and the spread of infectious diseases: Application to influenza in France. *PLOS ONE* **9**, e83002 (2014). Medline doi:10.1371/journal.pone.0083002

7. N. Eagle, M. Macy, R. Claxton, Network diversity and economic development. *Science* **328**, 1029–1031 (2010). Medline doi:10.1126/science.1186605

8. V. Padmanabhan, R. Ramjee, P. Mohan, U.S. Patent 8,423,255 (2013).

9. G. Boulton, Open your minds and share your results. *Nature* **486**, 441 (2012).

10. M. McNutt, Journals unite for reproducibility. *Science* **346**, 679 (2014). Medline doi:10.1126/science.aaa1724

11. T. Bloom, “Data access for the open access literature: PLOS’s data policy” (2013); www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy.

12. K. Burns, “In US cities, open data is not just nice to have; it’s the norm” *The Guardian*, 21 October 2013; www.theguardian.com/local-government-network/2013/oct/21/open-data-us-san-francisco.

13. Massachusetts Bay Transportation Authority, “Real-time commuter rail data” (2010); www.mbta.com/rider_tools/developers/default.asp?id=21899.

14. Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, V. D. Blondel, D4D-Senegal: The second mobile phone data for development challenge. (2014); http://arxiv.org/abs/1407.4885.

15. V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, C. Ziemlicki, Data for Development: The D4D challenge on mobile phone data. (2012); http://arxiv.org/abs/1210.0137.

16. P. Mutchler, “MetaPhone: The sensitivity of telephone metadata” (2014); http://webpolicy.org/2014/03/12/metaphone-the-sensitivity-of-telephone-metadata/.

17. Y.-A. de Montjoye, J. Quoidbach, F. Robic, A. Pentland, Predicting personality using novel mobile phone-based metrics. in *Proc. SBP* (Springer, Berlin, Heidelberg, 2013), pp. 48–55.

18. P. M. Schwartz, D. J. Solove, Reconciling personal information in the United States and European Union. *Calif. Law Rev.* **102**, 877–916 (2014).

19. Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.* **3**, 1376 (2013). Medline doi:10.1038/srep01376

20. A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets. in *IEEE Symposium on Security and Privacy*, Oakland, CA, 18 to 22 May 2008 (IEEE, New York, 2008), pp. 111–125.

21. A. C. Solomon, R. Hill, E. Janssen, S. A. Sanders, J. R. Heiman, Uniqueness and how it impacts privacy in health-related social science datasets. in *Proc. IHI* (Association for Computing Machinery, New York, 2012), pp. 523–532.

22. L. Sweeney, k-Anonymity: A model for protecting privacy. *Int. J. Unc. Fuzz. Knowl. Based Syst.* **10**, 557–570 (2002). doi:10.1142/S0218488502001648

23. 2013 Federal Reserve payments study (2013); www.frbservices.org/files/communications/pdf/research/2013_payments_study_summary.pdf.

24. eMarketer, "US mobile payments to top $ 1 billion in 2013" (2013); www.emarketer.com/Article/US-Mobile-Payments-Top-1-Billion-2013/1010035.

25. "The trust advantage: How to win with big data" (2013); www.bcgperspectives.com/content/articles/information_technology_strategy_consumer_products_trust_advantage_win_big_data/.

26. C.-L. Huang, M.-C. Chen, C.-J. Wang, Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl.* **33**, 847–856 (2007). doi:10.1016/j.eswa.2006.07.007

27. S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, Data mining for credit card fraud: A comparative study. *Decis. Support Syst.* **50**, 602–613 (2011). doi:10.1016/j.dss.2010.08.008

28. C. Krumme, A. Llorente, M. Cebrian, A. S. Pentland, E. Moro, The predictability of consumer visitation patterns. *Sci. Rep.* **3**, 1645 (2013). Medline doi:10.1038/srep01645

29. Materials and methods are available as supplementary materials on *Science* Online.

30. European Commission, "General data protection regulation" (2012); http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.

31. Deutsche Telekom, "Guiding principle big data" (2014); www.telekom.com/static/-/205808/1/guiding-principles-big-data-si.

32. Y.-A. de Montjoye, J. Kendall, C. Kerry, Enabling Humanitarian Use of Mobile Phone Data. *Brookings Issues in Technology Innovation Series* (Brookings Institution, Washington, DC, 2014), vol. 26.

33. Y.-A. de Montjoye, S.S. Wang, A.S. Pentland, On the trusted use of large-scale personal data. *IEEE Data Eng. Bull.* **35**, 5–8 (2012).

34. C. Dwork, Differential privacy. in *Automata, Languages and Programming* (Lecture Notes in Computer Science Series, Springer, Berlin, Heidelberg, 2006), vol. 4052, pp. 1–12.

35. D. DeSieno, Adding a conscience to competitive learning. *in IEEE International Conference on Neural Networks*, San Diego, CA, 24 to 27 July 1988 (IEEE, New York, 1988), pp. 117–124.